AgriGPT: a Large Language Model Ecosystem for Agriculture

Bo Yang[1], Yu Zhang[1], Lanfei Feng[2], Yunkui Chen[2], Jianyu Zhang[1], Xiao Xu[1], Nueraili Aierken[1], Yurui Li[1], Yuxuan Chen[1], Guijun Yang[3], Yong He[4], Runhe Huang[5], Shijian Li[1*]

[1]College of Computer Science and Technology, Zhejiang University, China
[2]College of Software Technology, Zhejiang University, China
[3]School of Geological Engineering and Geomatics, Chang'an University, China
[4]College of Biosystems Engineering and Food Science, Zhejiang University, China
[5]Faculty of Computer and Information Sciences, Hosei University, Japan
{boyang30, 22421173, jianyu.zhang, 3200105334, nureli, liyr, yuxuan_chen, yhe, shijianli}@zju.edu.cn,
{22451116, 22351048}@zju.edu.cn, yanggj@chd.edu.cn, rhuang@hosei.ac.jp

Abstract

Despite the rapid progress of Large Language Models (LLMs), their application in agriculture remains limited due to the lack of domain-specific models, curated datasets, and robust evaluation frameworks. To address these challenges, we propose **AgriGPT**, a domain-specialized LLM ecosystem for agricultural usage. At its core, we design a multi-agent scalable data engine that systematically compiles credible data sources into **Agri-342K**, a high-quality, standardized question-answer (QA) dataset. Trained on this dataset, AgriGPT supports a broad range of agricultural stakeholders, from practitioners to policy-makers. To enhance factual grounding, we employ **Tri-RAG**, a three-channel Retrieval-Augmented Generation framework combining dense retrieval, sparse retrieval,and multi-hop knowledge graph reasoning, thereby improving the LLM reasoning reliability. For comprehensive evaluation, we introduce **AgriBench-13K**, a benchmark suite comprising 13 tasks with varying types and complexities. Experiments demonstrate that AgriGPT significantly outperforms general-purpose LLMs on both domain adaptation and reasoning. Beyond the model itself, AgriGPT represents a modular and extensible LLM ecosystem for agriculture, comprising structured data construction, retrieval-enhanced generation, and domain-specific evaluation. This work provides a generalizable spectrum for developing scientific and industry specialized LLMs. All models, datasets, and code will be released to empower agricultural communities, especially in underserved regions, and promote open, impactful research.

Introduction

Agriculture is central to global food security and sustainability (Short Jr et al. 2023; Di Terlizzi 2016; Swaminathan 2001; Rozenstein et al. 2024; Kamilaris and Prenafeta-Boldu 2018; Wolfert et al. 2017; Liakos et al. 2018; Foley et al. 2011; Clapp 2020), yet remains underrepresented in AI research. Effective decisions rely on integrating diverse knowledge from crop science to market signals. Addressing this gap is critical, especially for smallholder farmers facing knowledge constraints (Godfray et al. 2010; Altieri 2009; Rockström et al. 2017; Calicioglu et al. 2019; Fan and Rue 2020; Klerkx and Rose 2020; Vanlauwe et al. 2014; Rivera and Sulaiman 2009).

Recent progress in LLMs (Chen et al. 2024; Jiang et al. 2023; Zha et al. 2023; Fang et al. 2023; Zhao et al. 2023; Achiam et al. 2023) shows promise in enabling multimodal reasoning, but their direct application in agriculture is limited by sparse terminology coverage, lack of retrieval grounding, and domain-specific reasoning gaps (Rezayi et al. 2022; Yang et al. 2024). Agricultural data are fragmented, unstructured, and domain experts are scarce (Wigboldus et al. 2016). Moreover, successful deployment requires not only accuracy, but local relevance, interpretability, and infrastructure-awareness (Ghanem 2015).

Several domain-specific LLMs exist-e.g.,AgroGPT (Awais et al. 2025), AgroLLM (Samuel et al. 2025), AgriLLM (Didwania et al. 2024), and AgriBERT (Rezayi et al. 2022), as well as task-specific applications such as LLM-based QA systems (He 2024) and plant disease detection models (Zhao et al. 2024).In other domains, notable examples include BioGPT (Luo et al. 2022), MedPaLM (Tu et al. 2024), LegalGPT (Shi et al. 2024). However, most of these models lack generative reasoning capabilities, depend on complex pipelines, or are not open-source. Tools like ShizishanGPT (Yang et al. 2024) rely on complex retrieval pipelines. In contrast, AgriGPT is a unified, open ecosystem integrating scalable instruction curation and reasoning.

Unlike prior works such as MAG-v (Sengupta et al. 2024), BioInstruct (Tran et al. 2024), and Self-Instruct (Wang et al. 2022), we construct **Agri-342K**, a large instruction dataset using our multi-agent **data engine** generation. Our retrieval module (**Tri-RAG**) combines dense, sparse, multi-

*Corresponding author

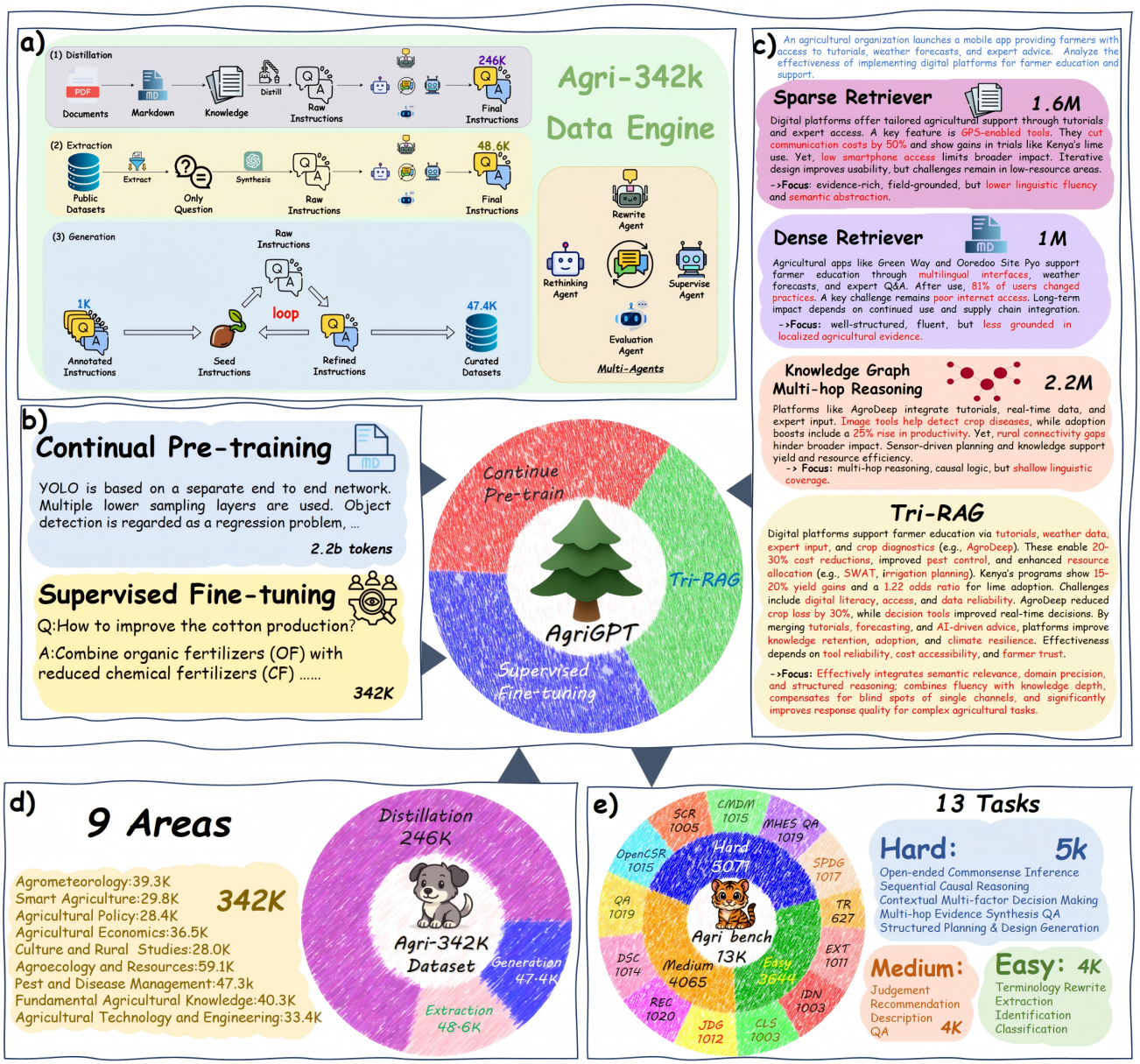

Figure 1: The AgriGPT Ecosystem: a). AgriGPT Data Engine: the 3 pipelines to construct Agri-342K dataset b). illustrating the model training workflow (continual pretraining and supervised fine-tuning) c). Tri-RAG inference and ablation: highlighting multi-path gains over single-path baselines d). the Agri-342K dataset with a broad topic spectrum e). the AgriBench-13K benchmark design

hop knowledge graph reasoning strategies (Lewis et al. 2020; Xiong et al. 2020; Rackauckas 2024; Sanmartin 2024; Peng et al. 2024; Arslan et al. 2024; Jin et al. 2024) to address factuality and reasoning gaps. We also introduce **AgriBench-13K**, a benchmark covering factual QA, diagnostic reasoning, , multi-hop reasoning and several tasks.

By releasing all code, data, and benchmarks, we aim to lower barriers to agricultural AI deployment. Grounded in global development and AI-for-social-good discourse (Vinuesa et al. 2020; Crawford 2021; Shi, Wang, and Fang 2020), AgriGPT supports equitable access to intelligent tools for farmers, and lays a foundation for scalable, socially impactful LLM applications in agriculture.

To sum up, our contributions are as follows:

- **Agri-342K dataset:** we develop a multi-agent data engine that enables scalable, high-quality curation of agricultural knowledge to the creation of 342K instruction data, along with a multilingual version to support cross-

lingual capabilities.

- **AgriBench-13K Benchmark Suite:** A multi-task, multi-level benchmark for agricultural LLMs, including Mini-AgriBench600 and its multilingual variant for lightweight and cross-lingual evaluation.
- **AgriGPT Ecosystem:** We propose the first comprehensive LLM ecosystem for agriculture and open-source its model, dataset, and benchmark, while providing a scalable and transferable framework—via our multi-agent engine and Tri-RAG—for building domain-specific LLMs in other real-world fields.

AgriGPT

In this section, we present the AgriGPT ecosystem (Figure 1). Our methods are derived as follows: we first introduce our data engine, a multi-agent data creation pipeline that constructs the Agri-342K instruction dataset and AgriBench-13K. Then, we describe the continual pre-training stage for domain adaptation. This is followed by supervised fine-tuning to align the model with task-specific objectives. Next, we present Tri-RAG, a retrieval-augmented inference module that integrates dense retrieval, sparse retrieval,and multi-hop knowledge graph reasoning to enhance factual accuracy and reasoning. Finally, we detail the development of AgriBench-13K, a multi-task benchmark for evaluating model performance across diverse agricultural tasks and difficulty levels.

Data Engine

We define the scope of agricultural knowledge by organizing it into 9 major thematic domains, each representing a core area of agricultural science and practice. To ensure comprehensive coverage within each domain, we manually curated a list of over 600 sub-area keywords, which were used to guide document retrieval, filtering, and categorization. Based on these keywords, we collected a wide range of credible, domain-relevant documents from research papers, technical manuals, and textbooks. Thematic distribution of these documents across the 9 domains is shown in Table 1.

As illustrated in Figure 1(a), we develop a modular and scalable data creation engine that constructs instructions. The data engine consists of 3 pipelines to systematically aggregate variety of data sources into unified data format. For publicly available data, we use distillation to process credible documents as raw instructions. For existing public question datasets (SivaResearch 2024; KisanVaani 2024), we employ general domain LLM (DeepSeek-R1-671B (Guo et al. 2025)) to generate raw answers. For scaling up data synthesis, we initiate a loop by 1K human verified instructions as seed to continuously generate new raw instructions, while accepting a portion of curated data to enrich the seed and improve synthetic quality.

For all the raw instructions from 3 pipelines, we design 4 collaborative agents powered by DeepSeek-R1-671B to finalize them into logical, diverse and factual grounded instructions: As illustrated in Figure 2. First, the **Rethinking Agent** revisits each QA pair by simulating alternative reasoning paths and exploring different semantic perspectives. This process helps uncover logical gaps, improve coverage, and enhance the diversity and robustness of the content. Next, the **Rewrite Agent** ensures stylistic coherence and linguistic normalization by paraphrasing the text, standardizing format and tone, and enforcing the use of domain-specific terminology, all while aligning with established instructional prompting practices. The **Supervise Agent** then acts as a semantic validator, checking alignment between the question and its source context, verifying factual correctness, and filtering hallucinated or off-topic content. Finally, the **Evaluation Agent** applies both rule-based and model-driven scoring metrics to assess each QA pair across dimensions such as coherence, informativeness, and accuracy. Based on these evaluations, only the highest-quality samples are retained, ensuring the overall reliability and instructional value of the dataset.

Areas	Num
Fundamental Agri Knowledge	11306
Pest and Disease Management	17856
Agroecology and Natural Resources	24947
Agri Technology and Engineering	21000
Smart Agri, AI & Computing	21062
Agri Economics	17741
Meteorology, Remote Sensing	29884
Agricultural Policy and Governance	17914
Life, Culture and Rural Studies	20304
Total	**182014**

Table 1: Thematic distribution of agricultural documents

The statistics of 3 pipelines are as follows:

- **Distillation:** We process approximately 182k research papers and 591 foundational books using MinerU to extract clean markdown-formatted content. All documents are first passed through a rule-based preprocessing pipeline, including structure-aware parsing, noise removal (e.g., boilerplate, references, page numbers), and keyword-based filtering to ensure a high signal-to-noise ratio. To further improve content relevance, we apply BM25 filtering guided by our 600+ sub-area taxonomy. The retained passages are then processed through our *collaborative multi-agent framework*, which orchestrates prompt construction, content paraphrasing, and iterative quality control. This distillation process transforms raw literature into high-quality, domain-specific QA pairs that reflect scientific depth and terminological precision.
- **Extraction:** We collect agriculture related questions from public QA datasets and encyclopedic sources. Since many entries lack complete answers or follow inconsistent formats, we reformat them into instruction-answer pairs using prompt-based rewriting. These samples are then passed through the *collaborative multi-agent framework*, where the answers are regenerated to align with the agricultural language conventions and evaluated for coherence, correctness and topical alignment.

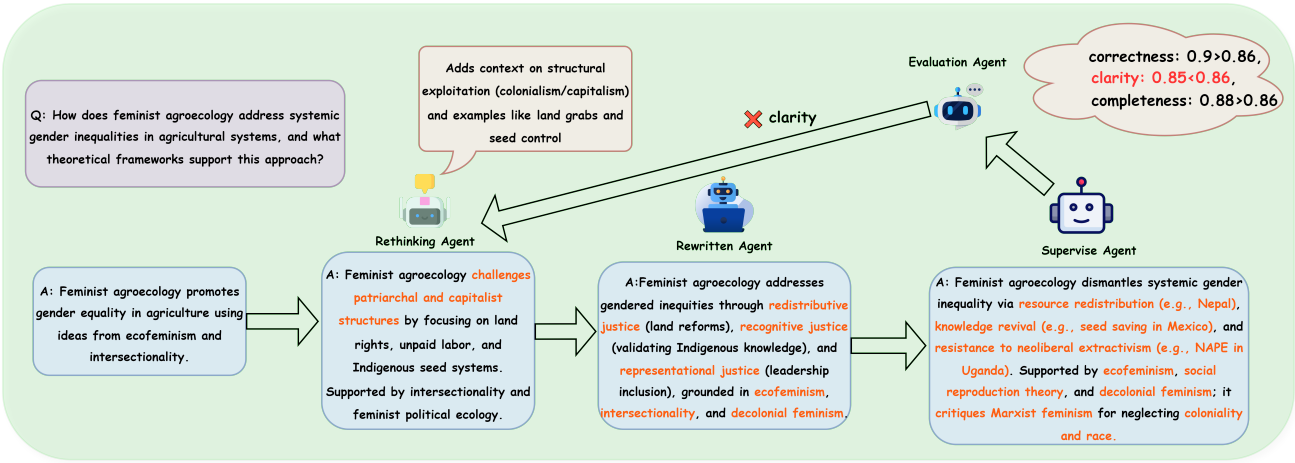

Figure 2: Workflow: Multi-Agent Framework for Ensuring Instruction Data Quality

- **Generation:** We manually compose a set of about 1K expert-written seed prompts covering diverse agricultural topics and task types. These prompts serve as anchors for iterative prompt expansion. Using the *collaborative multi-agent framework*, we generate rephrased, extended, and related question variants along with corresponding answers through multi-round prompting. The outputs are deduplicated, reviewed, and filtered to ensure factual accuracy and diversity. We also pay special attention to expert agreement and noise control during annotation to ensure high-quality supervision, resulting in a robust supplemental QA set.

The full pipeline yields over 342K reasoning-augmented QA pairs covering almost all areas of agriculture. Our data engine offers a reusable framework for constructing domain-specific instruction datasets at scale.

Continual Pretraining Stage

As shown in Figure 1 b), before fine-tuning on Agri-342K, we perform a continual pretraining stage on Qwen3-8B (Yang et al. 2025). We adopt a LoRA-based (Hu et al. 2022) approach to learn agricultural terminology and linguistic patterns from the corpus while maintaining its general language capabilities. As an extension of the model's original pretraining, this process helps AgriGPT absorb specialized vocabulary (e.g., crop species names, disease terminology) and factual knowledge from agricultural literature. The continual pretraining phase mitigate catastrophic forgetting of general knowledge and effectively adapt to the agricultural domain, laying a solid foundation for subsequent supervised fine-tuning.

Supervised Fine-Tuning Stage

As shown in Figure 1 b), after continual pretraining, we perform Supervised Fine-Tuning (SFT) on the Agri-342K dataset created by the data engine. In this stage, the model learns to follow a question and generate a correct answer in our dataset format. We fine-tune the model's parameters (including the LoRA adapter weights and, optionally, the top layers of the base model) using a standard language modeling objective to maximize the probability of the correct answer given the question with cross entropy loss. The fine-tuning process exposes the model to a wide variety of agriculture queries: from simple factual questions to complex explanatory prompts—and their corresponding answers. As a result, AgriGPT learns to produce high-quality answers that are both informative and domain-appropriate. The SFT stage solidifies the model's ability to perform as an agricultural assistant, aligning its generation style with the QA format and content of our curated dataset.

Tri-RAG Stage

To enhance AgriGPT's ability to handle complex agricultural queries, we propose a three channel Retrieval-Augmented Generation (Tri-RAG) framework that incorporates three complementary information sources at inference time. The design centers on combining dense semantic matching, domain-focused filtering, and structured reasoning to provide rich and diverse external context for generation.

The first channel builds a dense semantic index over the credible research papers and agricultural textbooks from data sources, allowing the model to retrieve document that are semantically aligned with the query. The second channel applies a BM25-based sparse retrieval strategy to extract high-relevance agricultural fragments from the same corpus, ensuring that the retrieved content is both targeted and domain-specific. The third channel further processes the BM25-filtered content by extracting approximately 2 million factual triples in the form of (entity, relation, entity), which are then used to construct a knowledge graph for multi-hop graph-based reasoning.

Each retrieval channel contributes uniquely: the dense retriever offers semantically fluent matches, the sparse retriever ensures domain-specific precision, and the graph-

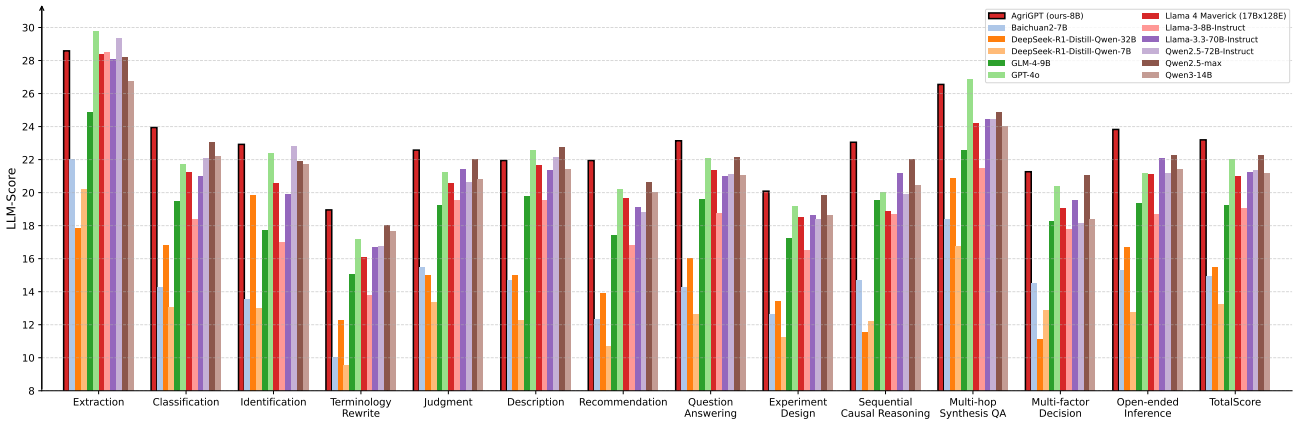

Figure 3: LLM-Based Evaluation of AgriGPT and Other Models across 13 Tasks of AgriBench-13K and Total Score

based retriever enables structured, multi-hop reasoning. During inference, outputs from all three channels are merged and re-ranked using a composite relevance scoring function that accounts for both retrieval confidence and content diversity. This process filters out redundant or low-quality information, ensuring that only the most informative and complementary evidence is selected. The top-ranked passages and/or factual triples are then integrated with the original query to construct an augmented prompt, which is fed into AgriGPT. Leveraging its pretraining on agricultural corpora and instruction fine-tuning, the model can effectively utilize this contextual information during generation.

As shown in Figure 1 c), We clearly demonstrate the performance differences between using a single-channel RAG and the Tri-RAG. By combining dense semantic alignment, domain-specific retrieval, and graph-based reasoning, the Tri-RAG framework empowers AgriGPT to produce responses that are factually grounded, contextually relevant, and logically coherent. This multi-channel design significantly enhances the model's performance on complex, knowledge-intensive agricultural tasks that require both precise retrieval and multi-step reasoning.

Easy Tasks	**Num**
Extraction	1,011
Classification	1003
Identification	1003
Terminology Rewrite	627
Medium Tasks	**Num**
Judgment	1,012
Description	1,014
Recommendation	1,020
Question Answering	1,019
Hard Tasks	**Num**
Experiment Design	1,017
Sequential Causal Reasoning	1,005
Multi-hop Evidence Synthesis QA	1,019
Contextual Multi-factor Decision	1,015
Open-ended Commonsense Inference	1,015

Table 2: Task distribution of AgriBench-13K

AgriBench-13K

As shown in Figure 1 e), to systematically assess the capabilities of LLMs in agriculture, we present AgriBench-13K, a comprehensive benchmark specifically designed for evaluating LLMs in agricultural contexts. It comprises 13 representative task types, these tasks reflect a wide range of language understanding and reasoning challenges encountered in real-world agricultural scenarios. The distribution of task types is summarized in Table 2.

The benchmark construction begins with domain experts defining 13 distinct task categories. Based on 9 major agricultural domains and over 600 sub-area labels, a total of 585 seed prompts are created, each corresponding to a core problem instance within a specific task and sub-area combination. These seeds are expanded using our data engine, which iteratively generates large-scale candidate question–answer pairs through cyclic sampling and multi-round refinement. All samples are subsequently de-duplicated and manually reviewed to ensure quality and consistency. To ensure fairness, we strictly separate the benchmark from the training data (Agri-342K) and apply similarity-based filtering to avoid data leakage.

The resulting AgriBench-13K benchmark includes 12,780 high-quality samples, with broad topical and task coverage. For efficient evaluation, we also uniformly sample a lightweight subset, Mini-AgriBench600. As the first large-scale standardized benchmark in the agricultural domain, it offers a unified framework for evaluating the performance of agricultural LLMs and provides a foundation for advancing model development, training strategies, and system-level innovations in agriculture.

Model	BLEU	Meteor	Rouge-1	Rouge-2	Rouge-L	LLM-Score
Baichuan2-7B	1.57	14.13	20.55	4.41	19.64	14.96
DeepSeek-R1-Distill-Qwen-7B	1.91	13.60	20.22	4.59	19.34	13.21
GLM-4-9B	9.46	31.23	27.55	6.33	26.31	19.22
Meta-Llama-3-8B-Instruct	7.81	28.56	25.43	5.78	24.28	19.05
DeepSeek-R1-Distill-Qwen-32B	3.06	15.03	21.52	5.66	20.50	15.49
Llama-3.3-70B-Instruct	8.25	30.07	26.25	6.31	25.04	21.23
Qwen3-14B	7.63	25.27	<u>29.20</u>	<u>8.20</u>	<u>27.90</u>	21.19
Qwen2.5-72B-Instruct	11.29	33.52	28.31	7.23	27.04	21.32
GPT-4o	8.41	27.97	27.58	6.97	26.27	22.02
Llama 4 Maverick (17Bx128E)	7.37	28.37	26.76	6.60	25.55	20.98
Qwen2.5-max	<u>12.38</u>	<u>38.14</u>	28.70	7.10	27.30	<u>22.27</u>
AgriGPT (ours)	**16.52**	**44.06**	**31.60**	**9.63**	**30.32**	**23.20**

Correctness	Match ability	Fluency	Coherence	Relevance	Logical Consistency	Completeness
2.26	2.06	2.37	2.02	2.19	2.03	2.02
1.99	1.78	2.14	1.86	1.97	1.78	1.69
3.01	2.60	2.94	2.61	2.78	2.64	2.64
2.93	2.59	2.91	2.62	2.79	2.62	2.59
2.37	2.18	2.38	2.12	2.26	2.10	2.08
3.27	2.90	3.22	2.88	3.10	2.91	2.95
3.24	2.76	3.22	2.93	<u>3.23</u>	2.90	2.90
3.30	2.94	3.19	2.93	3.11	2.94	2.91
3.41	3.05	<u>3.27</u>	3.03	3.19	<u>3.05</u>	3.02
3.26	2.89	3.12	2.87	3.06	2.91	2.89
<u>3.53</u>	<u>3.08</u>	3.27	<u>3.07</u>	3.22	3.04	<u>3.07</u>
3.59	**3.16**	**3.44**	**3.19**	**3.36**	**3.21**	**3.25**

Table 3: Comparison of AgriGPT with general LLMs. **Bold** and <u>underlined</u> indicate best and second-best performance.

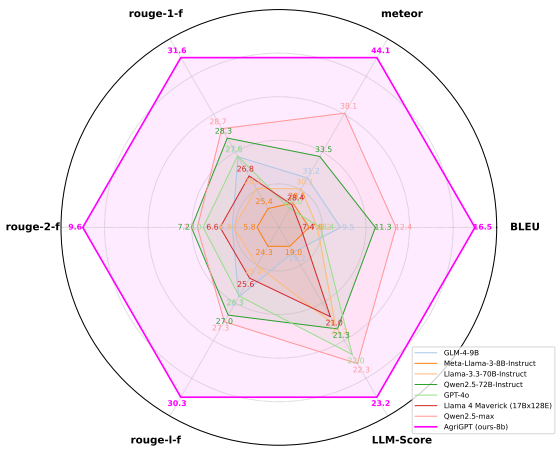

Figure 4: Performance comparison of AgriGPT and other models on AgriBench-13K

Result

Comparative Experiments

As illustrated in Table 3 and Figure 4, We evaluate AgriGPT against eleven representative LLMs: Baichuan2-7B (Yang et al. 2023), DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI 2025), GLM-4-9B (GLM et al. 2024), Meta-Llama-3-8B-Instruct (Dubey et al. 2024), DeepSeek-R1-Distill-Qwen-32B (DeepSeek-AI 2025), Llama-3.3-70B-Instruct (AI 2024a), Qwen3-14B (Yang et al. 2025), Qwen2.5-72B-Instruct (Qwen Team 2024b), GPT-4o (Achiam et al. 2023), Llama 4 Maverick (17B×128E) (AI 2024b), and Qwen2.5-max (Qwen Team 2024a). Overall and task-specific results are presented detailed in Appendix.

To the best of our knowledge, there are currently may not have open-source domain-specific LLMs for agriculture such as AgroGPT(Awais et al. 2025), AgroLLM (Samuel et al. 2025), AgriLLM (Didwania et al. 2024), AgriBERT (Rezayi et al.2022). AgriGPT is the first open-source model in this vertical domain, and thus we compare it only with general-purpose LLMs. Overall and task-specific results are presented in detail in the Appendix.

We adopt a dual evaluation strategy combining automatic

metrics and LLM-based scoring. Traditional metrics such as BLEU (Papineni et al. 2002), METEOR (Denkowski and Lavie 2014), and ROUGE (Lin 2004) quantify surface-level accuracy, while Qwen2.5-72B serves as an expert evaluator to score outputs across seven qualitative dimensions: Correctness, Match Ability, Fluency, Coherence, Relevance, Logical Consistency, and Completeness. Each dimension is scored from 0 (failure) to 5 (excellent), accompanied by a confidence score (0 to 1). The final LLM-Score is computed as a confidence-weighted average, improving robustness by prioritizing high-certainty judgments.

This hybrid approach ensures a balanced assessment of both lexical fidelity and deeper semantic quality. It demonstrates that AgriGPT not only delivers accurate responses but also generates coherent, logically consistent, and contextually relevant content—critical for real-world agricultural applications.

From the results, we observe the following. (1) **AgriGPT consistently achieves top scores** across both automatic and LLM-based evaluation strategies, highlighting its overall effectiveness. (2) **it also attains best ratings across all LLM-evaluated dimensions**, accompanied by high confidence scores, indicating both strong generation quality and reliable semantic alignment. (3) **the agreement between automatic metrics and LLM-based evaluations** reflects the balanced design of AgriGPT, which performs well at both surface-level accuracy and deeper reasoning. (4) **despite its relatively compact size**, AgriGPT outperforms many larger closed-source commercial LLMs, striking a strong balance between performance and efficiency. Its lightweight design enables deployment in low-resource regions and on constrained devices, enhancing real-world and social impact.

Figure 3 shows the LLM evaluation results of AgriGPT and other purpose-specific models across all 13 tasks in AgriBench-13K. AgriGPT outperforms across the majority of tasks, with particularly strong results in multi-hop reasoning and sequential understanding, demonstrating its versatility and robustness in complex agricultural scenarios.

Generalization Evaluation

To ensure AgriGPT avoids domain overfitting, we evaluate its generalization by comparing performance against the base model on several public general-domain benchmarks, including MMLU(Hendrycks et al. 2020; Mathew, Karatzas, and Jawahar 2021) , ARC (Clark et al. 2018), and OpenBookQA (Mihaylov et al. 2018)—three widely-used benchmarks that span diverse knowledge areas such as science, humanities, and logical reasoning, and are standard for evaluating language model generalization. The results (Table 4) show that AgriGPT performs comparably to Qwen3-8B, with only marginal differences within 0.5 points across datasets. This demonstrates that our domain adaptation enhances agricultural capabilities without compromising general reasoning and language skills.

Multilingual Evaluation

To evaluate multilingual capabilities, we translated Agri-342K into multiple languages to construct Agri-342K-Multilingual and conducted instruction tuning on AgriGPT.

Model	MMLU	ARC	OpenBookQA
Qwen3-8B(base)	85.87%	97.56%	91.77%
AgriGPT(ours)	85.84%	97.49%	91.20%

Table 4: Generalization to public benchmarks

For efficient testing, we sampled 600 examples from AgriBench-13K to create Mini-AgriBench600, then translated it into multiple languages (Mini-AgriBench600-Multilingual). As shown in Table 5, the model achieves reasonable BLEU and Meteor scores on Chinese and Japanese, indicating effective transfer of instruction-following ability across languages.

Language	BLEU	Meteor
English	16.52	40.16
Chinese	17.80	48.42
Japanese	16.69	45.30

Table 5: Multilingual Evaluation

Ablation Study

We conduct an ablation study to assess the effects of domain-specific training and RAG. As shown in Table 6, both components individually improve BLEU and Meteor scores, while their combination yields the best performance, highlighting their complementary benefits.

Model	BLEU	Meteor
Qwen3-8B	12.53	41.51
Qwen3-8B + RAG	13.39	42.06
Qwen3-8B + domain training	16.29	43.76
Qwen3-8B + domain training + RAG	16.42	44.15

Table 6: Ablation Study of Domain training and RAG

Potential Social Impact and Limitations

AgriGPT has the potential to significantly enhance agricultural productivity and decision-making in underserved rural regions. By enabling intelligent question answering, policy support, and real-time analysis in local agricultural contexts, it empowers farmers, extension workers, and policymakers with accessible, domain-specific knowledge. We further demonstrate its deployability by achieving 44.15 token/s inference speed on a single RTX 4090 GPU, enabling cost-effective edge deployment.This supports scalable usage in low-resource settings, helping reduce knowledge inequality, promote sustainable practices, and improve food security outcome.

AgriGPT currently has three key limitations: it only supports text input without multimodal capabilities; its training data lacks diversity due to reliance on formal sources; and it does not explicitly handle regional dialects. Future work will focus on adding image and sensor inputs, incorporating

informal texts and farmer dialogues, and expanding dialect coverage to improve real-world applicability.

Conclusion

We present AgriGPT, a domain-specific LLM designed to support complex agricultural tasks. Starting with a scalable, collaborative multi-agent data engine and a Tri-RAG framework, we obtain a high-quality Agri-342K instruction dataset that captures diverse agricultural knowledge. To systematically assess model performance, we introduce AgriBench-13K Suite, a comprehensive benchmark covering a wide range of agricultural tasks and difficulty levels. AgriGPT also retains strong performance on general-domain benchmarks and demonstrates effective multilingual transfer, reinforcing its robustness beyond agriculture-specific contexts. Together, the model, data engine, dataset, and benchmark form a coherent LLM ecosystem that not only advances agricultural AI research but also enables equitable and practical deployment in real-world, low-resource farming communities—highlighting its broader social impact and its potential to foster inclusive digital transformation in agriculture.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI, M. 2024a. Llama 3.3-70B-Instruct. https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct. Accessed: 2025-05-20.

AI, M. 2024b. LLaMA 4: Open Foundation and Instruction Models. Accessed: 2025-05-20.

Altieri, M. A. 2009. Agroecology, small farms, and food sovereignty. *Monthly review*, 61(3): 102–113.

Arslan, M.; Ghanem, H.; Munawar, S.; and Cruz, C. 2024. A Survey on RAG with LLMs. *Procedia Computer Science*, 246: 3781–3790.

Awais, M.; Alharthi, A. H. S. A.; Kumar, A.; Cholakkal, H.; and Anwer, R. M. 2025. Agrogpt: Efficient agricultural vision-language model with expert tuning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 5687–5696. IEEE.

Calicioglu, O.; Flammini, A.; Bracco, S.; Bellù, L.; and Sims, R. 2019. The future challenges of food and agriculture: An integrated analysis of trends and solutions. *Sustainability*, 11(1): 222.

Chen, Z. Z.; Ma, J.; Zhang, X.; Hao, N.; Yan, A.; Nourbakhsh, A.; Yang, X.; McAuley, J.; Petzold, L.; and Wang, W. Y. 2024. A survey on large language models for critical societal domains: Finance, healthcare, and law. *arXiv preprint arXiv:2405.01769*.

Clapp, J. 2020. Food security and nutrition: building a global narrative towards 2030.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv:1803.05457v1*.

Crawford, K. 2021. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Denkowski, M.; and Lavie, A. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, 376–380.

Di Terlizzi, B. 2016. Enhancing knowledge for food security. *Mediterra 2016. Zero Waste in the Mediterranean. Natural Resources, Food and Knowledge/International Centre for Advanced Mediterranean Agronomic Studies (CIHEAM) and Food and Agriculture Organization of the United Nations (FAO)–Paris: Presses de Sciences Po, 2016.*, 363.

Didwania, K.; Seth, P.; Kasliwal, A.; and Agarwal, A. 2024. AgriLLM: Harnessing Transformers for Farmer Queries. *arXiv preprint arXiv:2407.04721*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Ganapathy, R.; et al. 2024. The LLaMA 3 Herd of Models. *arXiv preprint arXiv:2407*.

Fan, S.; and Rue, C. 2020. The role of smallholder farms in a changing world. In *The role of smallholder farms in food and nutrition security*, 13–28. Springer International Publishing Cham.

Fang, Y.; Zhang, Q.; Zhang, N.; Chen, Z.; Zhuang, X.; Shao, X.; Fan, X.; and Chen, H. 2023. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nature Machine Intelligence*, 5(5): 542–553.

Foley, J. A.; Ramankutty, N.; Brauman, K. A.; Cassidy, E. S.; Gerber, J. S.; Johnston, M.; Mueller, N. D.; O'Connell, C.; Ray, D. K.; West, P. C.; et al. 2011. Solutions for a cultivated planet. *Nature*, 478(7369): 337–342.

Ghanem, H. 2015. Agriculture and rural development for inclusive growth and food security in Morocco. *Brookings Global Working Paper Series*.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Rojas, D.; Feng, G.; Zhao, H.; Lai, H.; Yu, H.; Wang, H.; Sun, J.; Zhang, J.; Cheng, J.; Gui, J.; Tang, J.; Zhang, J.; Li, J.; Zhao, L.; Wu, L.; Zhong, L.; Liu, M.; Huang, M.; Zhang, P.; Zheng, Q.; Lu, R.; Duan, S.; Zhang, S.; Cao, S.; Yang, S.; Tam, W. L.; Zhao, W.; Liu, X.; Xia, X.; Zhang, X.; Gu, X.; Lv, X.; Liu, X.; Liu, X.; Yang, X.; Song, X.; Zhang, X.; An, Y.; Xu, Y.; Niu, Y.; Yang, Y.; Li, Y.; Bai, Y.; Dong, Y.; Qi, Z.; Wang, Z.; Yang, Z.; Du, Z.; Hou, Z.; and Wang, Z. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793.

Godfray, H. C. J.; Beddington, J. R.; Crute, I. R.; Haddad, L.; Lawrence, D.; Muir, J. F.; Pretty, J.; Robinson, S.; Thomas, S. M.; and Toulmin, C. 2010. Food security: the challenge of feeding 9 billion people. *science*, 327(5967): 812–818.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1:

Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

He, X. 2024. Enhancing Agriculture QA Models Using Large Language Models. In *BIO Web of Conferences*, volume 142, 01005. EDP Sciences.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Jiang, J.; Zhou, K.; Dong, Z.; Ye, K.; Zhao, W. X.; and Wen, J.-R. 2023. Structgpt: A general framework for large language model to reason over structured data. *arXiv preprint arXiv:2305.09645*.

Jin, B.; Yoon, J.; Han, J.; and Arik, S. O. 2024. Long-context llms meet rag: Overcoming challenges for long inputs in rag. *arXiv preprint arXiv:2410.05983*.

Kamilaris, A.; and Prenafeta-Boldu, F. 2018. Deep learning in agri-culture: a survey. computers and electronics in agriculture 147: 70–90.

KisanVaani. 2024. Agriculture QA English Only. https://huggingface.co/datasets/KisanVaani/agriculture-qa-english-only. Accessed: 2024-05-20.

Klerkx, L.; and Rose, D. 2020. Dealing with the game-changing technologies of Agriculture 4.0: How do we manage diversity and responsibility in food system transition pathways? *Global Food Security*, 24: 100347.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474.

Liakos, K.; Busato, P.; Moshou, D.; Pearson, S.; and Bochits, D. 2018. Machine Learning in Agriculture: A Review. Sensors (Special Issue" Sensors in Agriculture 2018").

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Luo, R.; Sun, L.; Xia, Y.; Qin, T.; Zhang, S.; Poon, H.; and Liu, T.-Y. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6): bbac409.

Mathew, M.; Karatzas, D.; and Jawahar, C. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2200–2209.

Mihaylov, T.; Clark, P.; Khot, T.; and Sabharwal, A. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *EMNLP*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Peng, B.; Zhu, Y.; Liu, Y.; Bo, X.; Shi, H.; Hong, C.; Zhang, Y.; and Tang, S. 2024. Graph retrieval-augmented generation: A survey. *arXiv preprint arXiv:2408.08921*.

Qwen Team. 2024a. Qwen2.5-Max Announcement. https://qwenlm.github.io/blog/qwen2.5-max/.

Qwen Team. 2024b. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Rackauckas, Z. 2024. Rag-fusion: a new take on retrieval-augmented generation. *arXiv preprint arXiv:2402.03367*.

Rezayi, S.; Liu, Z.; Wu, Z.; Dhakal, C.; Ge, B.; Zhen, C.; Liu, T.; and Li, S. 2022. AgriBERT: Knowledge-Infused Agricultural Language Models for Matching Food and Nutrition. In *IJCAI*, 5150–5156.

Rivera, W. M.; and Sulaiman, V. R. 2009. Extension: object of reform, engine for innovation. *Outlook on agriculture*, 38(3): 267–273.

Rockström, J.; Williams, J.; Daily, G.; Noble, A.; Matthews, N.; Gordon, L.; Wetterstrand, H.; DeClerck, F.; Shah, M.; Steduto, P.; et al. 2017. Sustainable intensification of agriculture for human prosperity and global sustainability. *Ambio*, 46(1): 4–17.

Rozenstein, O.; Cohen, Y.; Alchanatis, V.; Behrendt, K.; Bonfil, D. J.; Eshel, G.; Harari, A.; Harris, W. E.; Klapp, I.; Laor, Y.; et al. 2024. Data-driven agriculture and sustainable farming: friends or foes? *Precision Agriculture*, 25(1): 520–531.

Samuel, D. J.; Skarga-Bandurova, I.; Sikolia, D.; and Awais, M. 2025. AgroLLM: Connecting Farmers and Agricultural Practices through Large Language Models for Enhanced Knowledge Transfer and Practical Application. *arXiv preprint arXiv:2503.04788*.

Sanmartin, D. 2024. Kg-rag: Bridging the gap between knowledge and creativity. *arXiv preprint arXiv:2405.12035*.

Sengupta, S.; Vashistha, H.; Curtis, K.; Mallipeddi, A.; Mathur, A.; Ross, J.; and Gou, L. 2024. Mag-v: A multi-agent framework for synthetic data generation and verification. *arXiv preprint arXiv:2412.04494*.

Shi, J.; Guo, Q.; Liao, Y.; and Liang, S. 2024. LegalGPT: Legal Chain of Thought for the Legal Large Language Model Multi-agent Framework. In *International Conference on Intelligent Computing*, 25–37. Springer.

Shi, Z. R.; Wang, C.; and Fang, F. 2020. Artificial intelligence for social good: A survey. *arXiv preprint arXiv:2001.01818*.

Short Jr, N. M.; Woodward-Greene, M. J.; Buser, M. D.; and Roberts, D. P. 2023. Scalable knowledge management to meet global 21st century challenges in agriculture. *Land*, 12(3): 588.

SivaResearch. 2024. Agri Dataset. https://huggingface.co/datasets/SivaResearch/Agri. Accessed: 2024-05-20.

Swaminathan, M. 2001. Food security and sustainable development. *Current Science*, 81(8): 948–954.

Tran, H.; Yang, Z.; Yao, Z.; and Yu, H. 2024. BioInstruct: instruction tuning of large language models for biomedical natural language processing. *Journal of the American Medical Informatics Association*, 31(9): 1821–1832.

Tu, T.; Azizi, S.; Driess, D.; Schaekermann, M.; Amin, M.; Chang, P.-C.; Carroll, A.; Lau, C.; Tanno, R.; Ktena, I.; et al. 2024. Towards generalist biomedical AI. *Nejm Ai*, 1(3): AIoa2300138.

Vanlauwe, B.; Coyne, D.; Gockowski, J.; Hauser, S.; Huising, J.; Masso, C.; Nziguheba, G.; Schut, M.; and Van Asten, P. 2014. Sustainable intensification and the African smallholder farmer. *Current Opinion in Environmental Sustainability*, 8(0): 15–22.

Vinuesa, R.; Azizpour, H.; Leite, I.; Balaam, M.; Dignum, V.; Domisch, S.; Felländer, A.; Langhans, S. D.; Tegmark, M.; and Fuso Nerini, F. 2020. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature communications*, 11(1): 233.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Wigboldus, S.; Klerkx, L.; Leeuwis, C.; Schut, M.; Muilerman, S.; and Jochemsen, H. 2016. Systemic perspectives on scaling agricultural innovations. A review. *Agronomy for sustainable development*, 36(3): 46.

Wolfert, S.; Ge, L.; Verdouw, C.; and Bogaardt, M.-J. 2017. Big data in smart farming–a review. *Agricultural systems*, 153: 69–80.

Xiong, W.; Li, X. L.; Iyer, S.; Du, J.; Lewis, P.; Wang, W. Y.; Mehdad, Y.; Yih, W.-t.; Riedel, S.; Kiela, D.; et al. 2020. Answering complex open-domain questions with multi-hop dense retrieval. *arXiv preprint arXiv:2009.12756*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Qiu, Z.; et al. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Yang, A.; Xiao, B.; Wang, B.; Zhang, B.; Bian, C.; Yin, C.; Lv, C.; Pan, D.; Wang, D.; Yan, D.; et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Yang, S.; Liu, Z.; Mayer, W.; Ding, N.; Wang, Y.; Huang, Y.; Wu, P.; Li, W.; Li, L.; Zhang, H.-Y.; et al. 2024. ShizishanGPT: An Agricultural Large Language Model Integrating Tools and Resources. In *International Conference on Web Information Systems Engineering*, 284–298. Springer.

Zha, L.; Zhou, J.; Li, L.; Wang, R.; Huang, Q.; Yang, S.; Yuan, J.; Su, C.; Li, X.; Su, A.; et al. 2023. Tablegpt: Towards unifying tables, nature language and commands into one gpt. *arXiv preprint arXiv:2307.08674*.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Zhao, X.; Chen, B.; Ji, M.; Wang, X.; Yan, Y.; Zhang, J.; Liu, S.; Ye, M.; and Lv, C. 2024. Implementation of large language models and agricultural knowledge graphs for efficient plant disease detection. *Agriculture*, 14(8): 1359.