AgriEval: A Comprehensive Chinese Agricultural Benchmark for Large Language Models

Lian Yan [1*], Haotian Wang[1*], Chen Tang[2*], Haifeng Liu[1], Tianyang Sun[1], Liangliang Liu[1], Yi Guan[1], Jingchi Jiang[1†]

[1]Harbin Institute of Technology
[2]Institute for Advanced Algorithms Research
{23B903008, 24S003142, 2021110688}@stu.hit.edu.cn, wanght1998@gmail.com
travistang@foxmail.com, {guanyi, jiangjingchi}@hit.edu.cn

Abstract

In the agricultural domain, the deployment of large language models (LLMs) is hindered by the lack of training data and evaluation benchmarks. To mitigate this issue, we propose AgriEval, the first comprehensive Chinese agricultural benchmark with three main characteristics: (1) Comprehensive Capability Evaluation. AgriEval covers six major agriculture categories and 29 subcategories within agriculture, addressing four core cognitive scenarios—memorization, understanding, inference, and generation. (2) High-Quality Data. The dataset is curated from university-level examinations and assignments, providing a natural and robust benchmark for assessing the capacity of LLMs to apply knowledge and make expert-like decisions. (3) Diverse Formats and Extensive Scale. AgriEval comprises 14,697 multiple-choice questions and 2,167 open-ended question-and-answer questions, establishing it as the most extensive agricultural benchmark available to date. We also present comprehensive experimental results over 51 open-source and commercial LLMs. The experimental results reveal that most existing LLMs struggle to achieve 60 percent accuracy, underscoring the developmental potential in agricultural LLMs. Additionally, we conduct extensive experiments to investigate factors influencing model performance and propose strategies for enhancement.

Code — https://github.com/YanPioneer/AgriEval/
Datasets — https://github.com/YanPioneer/AgriEval/
Extended version — https://arxiv.org/pdf/2507.21773

Introduction

The rapid development of large language models (LLMs) has enabled new applications in smart agriculture (Tzachor et al. 2023; Li et al. 2024c; Shahriar et al. 2025; Kuska, Wahabzada, and Paulus 2024), such as knowledge-based Q&A (Silva et al. 2023), cultivation planning (Peng et al. 2023), and plant science (Agathokleous et al. 2024; MacNish et al. 2025; Yang et al. 2024). However, agriculture is a highly specialized domain characterized by fragmented expertise, diverse subfields, and decision-making that requires reasoning over known conditions and intensive domain-specific knowledge. Open-domain LLMs, lacking sufficient agricultural pre-training and domain grounding, often produce factually incorrect or misleading outputs in this context.

Benchmark	Language	Domain	Data Type	Cognitive	Data Size
MMLU	English	General	MC	✗	–
CMMMU	Chinese	General	MC	✗	283
CMMLU	Chinese	General	MC	✗	348
C-Eval	Chinese	General	MC	✗	928
Agri-Eval	Chinese	Agriculture	MC&QA	✓	16,864

Table 1: AgriEval vs. existing benchmarks on agricultural Data. MC and QA denote multiple-choice and open-ended question answering formats, respectively.

To address these challenges, a dedicated benchmark is essential for systematically evaluating LLMs' capabilities in the agricultural domain. The proposal of such a benchmark not only reflects the performance and limitations of current models in agriculture but also provides valuable insights for the development and enhancement of agriculture-specific LLMs. Existing benchmarks (Huang et al. 2024; Li et al. 2024b; Zhang et al. 2024; Hendrycks et al. 2020) predominantly focus on general or semi-professional knowledge, with limited coverage of agricultural topics, as illustrated in Table 1. These studies, when considered as benchmarks for Agricultural AI, have two significant limitations: (1) *Extremely limited agriculture-related content* (typically $<$ 1.5% of total questions); and (2) *A lack of expert-level questions*, with most items focusing on basic knowledge (e.g., crop identification) rather than complex reasoning required for tasks such as precision disease diagnosis or pesticide formulation (Jiang et al. 2025). This dual deficiency, both in knowledge breadth and professional depth, renders current benchmarks inadequate for assessing LLMs' true competency in agricultural applications, where domain-specific knowledge and precise reasoning are critical for avoiding potentially serious real-world consequences.

In addition, benchmarks for Agricultural AI should introduce and account for more domain-specific challenges that extend beyond open-domain studies. For instance, regional diversity within the agricultural domain adds com-

*These authors contributed equally.
†Corresponding Author.

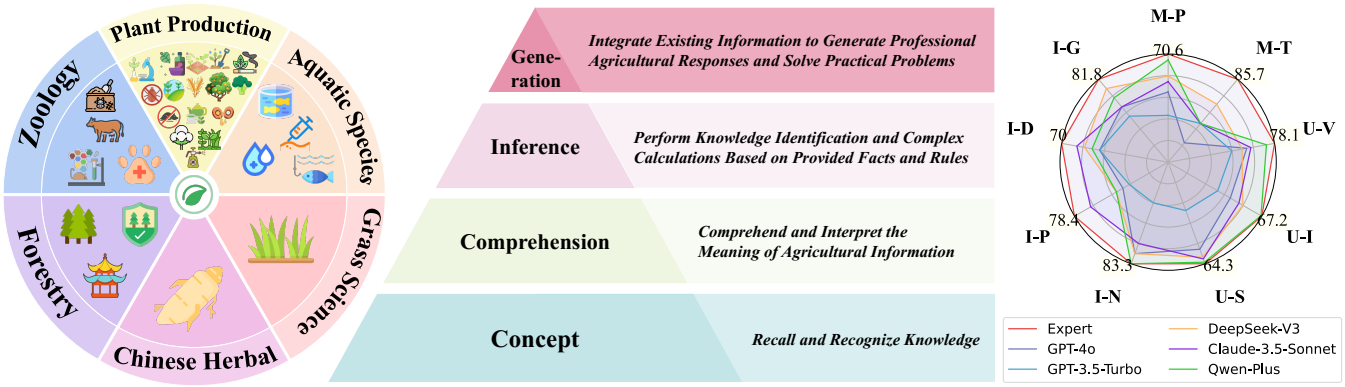

Figure 1: *Left*: Domains classification in AgriEval. *Middle*: Cognitive ability classification in AgriEval. *Right*: A brief overview of human experts and LLMs' performance on AgriEval.

plexity that tests the generalization capabilities of LLMs. In particular, Chinese agriculture poses unique difficulties due to its regional heterogeneity, ecological diversity, and cultural specificity (Jiang et al. 2025). Tasks such as pest control, crop breeding, and soil management are highly localized, while specialized subfields, such as traditional herbology and tea science, further broaden the domain scope. These factors should be incorporated into an Agricultural AI benchmark to comprehensively capture the breadth of agricultural knowledge and facilitate the fine-grained cognitive evaluation of LLMs.

To bridge the gap in LLM development and evaluation in the agricultural domain, we propose **AgriEval**, the first large-scale benchmark for cognitive assessment in real-world Chinese agricultural scenarios.[1] Developed under expert guidance, AgriEval covers six major categories and 29 subcategories (Figure 1 *Left*). To meet the high specialization demands of agricultural production, we collect 14,697 multiple-choice questions (including single-answer, multiple-answer, true/false) and 2,167 Q&A questions from both undergraduate and graduate-level professional exams. Inspired by Bloom's taxonomy (Seaman 2011; Li et al. 2024a) and real-world agricultural practices, AgriEval adopts a four-level cognitive framework—*Memorization*, *Understanding*, *Inference*, and *Generation*—further extended into 15 task-specific dimensions. This structure enables fine-grained evaluation of both knowledge breadth and reasoning depth in agricultural LLMs.

We conduct a comprehensive evaluation of 51 competitive LLMs on AgriEval, including nine commercial models and 42 open-source models spanning a wide range of architectures and parameter sizes. To assess their adaptability, we adopt multiple evaluation settings: zero-shot (Romera-Paredes and Torr 2015), few-shot (Snell, Swersky, and Zemel 2017), and chain-of-thought (CoT) (Wei et al. 2022; Chu et al. 2024). Furthermore, we explore option-shuffling, knowledge augmentation via retrieval-augmented generation (RAG) (Lewis et al. 2020), and analyze across cognitive levels and task types to probe models' internal reasoning patterns and external knowledge dependency.

Our experiments demonstrate that current LLMs struggle to reach the performance of a human primary expert, with even the most capable model, Qwen-Plus, achieving only 63.21% accuracy on AgriEval. Several key findings emerge from extensive experiments: (1) **Cognitive difficulty:** Performance varies significantly across cognitive levels, with numerical reasoning posing the greatest challenges. (2) **Positional sensitivity:** Most LLMs exhibit strong biases toward earlier options, performing poorly when correct answers appear in later positions. (3) **Scaling effects:** Models under 7B parameters average 34.15% accuracy; larger models generally perform better, though the trend is not strictly monotonic. (4) **Prompting strategies:** CoT boosts reasoning, particularly for complex tasks, while few-shot learning shows inconsistent gains. (5) **External knowledge integration:** RAG helps mitigate factual gaps in open-domain LLMs and improves performance on specialized tasks.

Related Work

Benchmarks play a critical role in evaluating model capabilities, akin to human-level examinations. Early benchmarks focused on task-specific objectives, such as machine translation (Bojar et al. 2014) and reading comprehension (Rajpurkar, Jia, and Liang 2018). With the emergence of LLMs, recent efforts have shifted toward evaluating general reasoning and embedded world knowledge (Zhang et al. 2024; Li et al. 2024a; Huang et al. 2024; Wang et al. 2024). MMLU (Hendrycks et al. 2020) consists of 15,908 multiple-choice questions across 57 subjects, spanning STEM and humanities, with varying levels of difficulty. Following this, more diverse benchmarks have gained traction. For instance, C-Eval (Huang et al. 2024) covers 52 Chinese-language disciplines, while CMMLU (Li et al. 2024b) extends this to 67 subjects. CMMMU (Zhang et al. 2024) broadens subject coverage across six core areas, including art, business, and science. MM-Vet (Yu et al. 2023) further explores model capabilities across tasks involving spatial reasoning, textual understanding, and mathematical problem-solving.

Despite their breadth, these benchmarks largely assess

[1]An English-translated version of these benchmarks has also been released and checked by GPT-4o.

non-specialized or semi-specialized knowledge. As pointed out by LexEval (Li et al. 2024a), they offer limited insight into domains demanding expert-level understanding, such as medicine, law, finance, and agriculture. To address this gap, several domain-specific benchmarks have emerged: CMD (Wang et al. 2023) for medicine, LexEval (Li et al. 2024a) for legal reasoning, and Golden Touchstone (Lee et al. 2023) for financial analysis. In the agricultural domain, existing efforts remain limited in scope and depth. AgriBench (Zhou and Ryo 2024) focuses primarily on multimodal perception tasks such as crop and scene segmentation, lacking fine-grained agricultural knowledge assessment. AgXQA (Kpodo, Kordjamshidi, and Nejadhashemi 2024) and BVL (Kästing and Hänig 2024) target narrow subdomains and small-scale datasets. AgriGPT (Yang et al. 2025), built upon synthetic LLM-generated Q&A, provides only coarse-grained classification without cognitive-level evaluation.

AgriEval

Task Overview

Motivation and Design Principles. Unlike previous benchmarks for LLMs, the benchmarks for Agricultural AI should incorporate and address more domain-specific challenges beyond open-domain studies, and they provide three primary resources: (1) A taxonomy of tasks and corresponding datasets that represent the capabilities an LLM should possess to function as an agricultural expert; (2) A systematic evaluation method to assess various types of LLMs regarding these capabilities; (3) Comprehensive experiments that demonstrate how mainstream LLMs perform on the agricultural benchmark, offering insights for developing an agriculture-specific LLM. In essence, AgriEval aims to provide a hierarchical cognitive taxonomy of agricultural tasks aligned with real-world decision-making practices. This design facilitates both model diagnostics and targeted improvements for practical deployment.

Domain Coverage. Following the human-expert benchmark paradigm (e.g., C-Eval (Huang et al. 2024)), AgriEval is curated from real examination questions designed for undergraduate and postgraduate students. With guidance from agricultural experts holding Ph.D. degrees in Agriculture within China's educational system, we align the domain taxonomy with China's official classification standards.[2] The benchmark spans six primary categories: *Plant Production* (PP), *Forestry* (Fore), *Grass Science* (GS), *Aquaculture* (Aqua), *Animal Science and Technology* (AST), and *Traditional Chinese Herbology* (TCH). These are further divided into 29 subfields, such as plant protection, smart agriculture, and tea science (see Figure 1 (*Left*)).

Cognitive Taxonomy. To assess both the breadth of knowledge and depth of reasoning in agricultural contexts, AgriEval introduces a four-level cognitive taxonomy inspired by Bloom's framework (Seaman 2011; Li et al. 2024a). The taxonomy consists of: (1) *Memorization*, which evaluates the recall of facts, terms, and procedures; (2) *Understanding*, which focuses on the ability to interpret, compare, and explain agricultural knowledge; (3) *Inference*, which assesses reasoning and problem-solving based on domain knowledge; and (4) *Generation*, which requires synthesizing information to produce professional, task-oriented responses. This hierarchical structure reflects the cognitive demands of real-world agricultural decision-making and supports fine-grained evaluation of LLMs. Complex tasks often span multiple levels, combining factual knowledge, reasoning, and domain-specific synthesis.

[2]https://www.gov.cn/zhengce/zhengceku/2020-12/30/content_5575377.htm

Level	Task	# Samples	Avg. Tokens
Memorization	Terminology Explanation (M-T)	125	101.85
	Fundamental Principles (M-P)	6,077	82.51
	Operational Rules (M-R)	116	89.7
	Production Management Essentials (M-E)	880	98.21
Understanding	Knowledge Verification (U-V)	1,961	45.75
	Type Identification (U-I)	2,253	80.3
	Key-Point Summarization (U-S)	1,324	103.32
Inference	Production Planning (I-P)	471	95.34
	Numerical Reasoning (I-N)	707	122.09
	Disease Diagnosis (I-D)	403	114.7
	Growth Status Analysis (I-S)	273	163.36
	Genetic Inference (I-G)	107	125.37
Generation	Knowledge-based Q&A (G-QA)	1,700	19.6
	Production Strategy Formulation (G-PS)	325	41.72
	Causal Analysis (G-CA)	142	22.92

Table 2: AgriEval cognitive ability data statistics.

Data Collection

We recruit two agricultural experts from partner agricultural university laboratories, both with advanced educational backgrounds, to collect domain-relevant examination materials from undergraduate and graduate-level assessments. Data sources include publicly available mock exams, graduate admission websites, and past exam materials publicly shared by students at top Chinese universities. All materials are originally in Word or PDF format. We collect over 500 documents and manually filter them based on difficulty, domain relevance, and alignment with real-world agricultural tasks, ultimately retaining 400 documents. The entire process spans approximately 1.0 months, with annotators compensated at 50 CNY per hour.

Data Annotation and Verification

To standardize the collected materials, all examination materials undergo a systematic digitization and structuring process. Source files in PDF are converted to Word documents using OCR and then parsed into a structured JSON format. The JSON schema contains attributes of the question, choices, answer, domain category, and cognitive category, and there are four question types, including single-choice, multiple-choice, true/false, and open-ended Q&A. For samples involving complex mathematical notation, expressions are manually converted into standard LaTeX format following the conventions of C-Eval (Huang et al. 2024) and MMLU (Hendrycks et al. 2020).

Each sample is initially categorized by agricultural experts using a custom annotation tool we developed. To ensure data quality, all entries are reviewed and corrected by expert annotators. To validate label consistency, we randomly sample 5% of the data and ask two experts to inde-

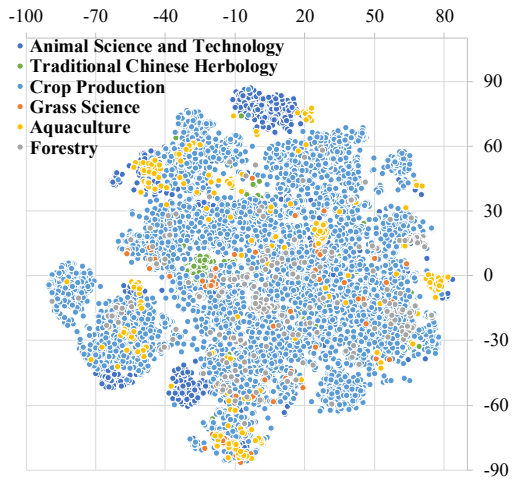

Figure 2: Question representation via BERT encoding and t-SNE dimensionality reduction.

pendently annotate the questions, answers, and labels. Inter-annotator agreement is computed, and disagreements are resolved through discussion and targeted retraining until consistency exceeds 90%. Only after reaching this threshold do we proceed with large-scale annotation. The Cohen's Kappa (Cohen 1960) score between expert annotators for category labeling is 0.85. Consistency for question/option correctness and answer correctness reaches 99.7%.

Data Difficulty Enhancement

To better assess the model's discriminative ability to distinguish between options and enhance the distractiveness of the choices, we follow the practice of C-Eval (Huang et al. 2024) and use GPT-4 to generate high-quality distractors. Each single-answer multiple-choice question is extended to include seven options, and all additional distractors are manually reviewed and validated by agricultural experts. This enhancement increases task difficulty while ensuring domain plausibility and consistency.

Data Statistics

Basic Statistics. AgriEval contains 14,697 multiple-choice questions and 2,167 open-ended Q&A items, covering 29 agricultural subfields and 15 cognitive categories. Each subfield includes at least 100 questions, while each cognitive category contains over 2,000 samples. The average question length is 76.92 tokens, and the average answer length for generation tasks is 467.30 tokens. Table 2 shows detailed distributions by cognitive level.

Semantic Diversity. To assess knowledge breadth, we visualize the semantic embedding space of AgriEval using BERT-based representations (Devlin et al. 2019). As shown in Figure 2, the embeddings demonstrate broad dispersion, indicating rich semantic coverage across domains. This suggests that AgriEval presents a diverse and challenging benchmark for LLMs.

Experiment

Experimental Setup

We conduct a systematic evaluation of 51 LLMs, comprising nine commercial and 42 open-source models that differ in architecture, parameter size, and language alignment. Open-source models are deployed on 4×NVIDIA H800 80GB GPUs via local inference, while commercial models are accessed through official APIs. All models are evaluated at a generation temperature of 1.0 and a maximum token length of 2048, with the results averaged over three independent runs to ensure stability.

To assess LLMs' knowledge understanding and reasoning capabilities in agricultural scenarios, we design multiple evaluation setups. We test three prompting strategies: **Zero-Shot Prompting**, where the model directly outputs an answer; **CoT Prompting**, which encourages step-by-step reasoning; and **Few-Shot Prompting**, which includes five in-context examples sampled from different subcategories. To examine the effect of external knowledge, we implement an **RAG** setup using Chinese Wikipedia as the retrieval corpus and evaluate it on a 1,000-sample subset. Additionally, to evaluate models' sensitivity to answer position, we randomly **shuffle multiple-choice options**, following prior work on positional bias in LLMs (Zheng et al. 2023; Du et al. 2024; Navigli, Conia, and Ross 2023).

For evaluation metrics, we use accuracy for all choice-based questions. In multi-answer questions, predictions are counted as correct only if all correct options are selected exactly. For open-ended Q&A tasks, we apply LLM-as-a-Judge (Gu et al. 2024) to measure generation quality.

Main Results

In this section, we conduct a comprehensive comparison of various LLMs on the AgriEval benchmark. To present the results more clearly, we highlight selected representative outcomes in Table 3 and Table 4. Based on these results, we summarize the key findings as follows:

AgriEval remains a highly challenging benchmark. LLMs achieve an average accuracy of 41.27% on AgriEval, with the vast majority of models failing to reach the 60% threshold. Even GPT-4o struggles with tasks requiring specialized agricultural knowledge, reflecting limited domain adaptation and reasoning capabilities. These results highlight both the difficulty of AgriEval in capturing real-world agricultural challenges and the substantial gap that remains for current LLMs to achieve expert-level performance in agricultural applications.

LLMs struggle with inference, revealing reasoning gaps. From a cognitive perspective, LLMs perform significantly worse on inference tasks—particularly those involving numerical reasoning and genetic inference—compared to memorization and understanding. This reflects a reliance on surface-level pattern recognition rather than grounded or compositional reasoning. For example, *solving a task such as design pump capacity and drainage modulus for a 3.8 km² polder under 200 mm rainfall and 2-day drainage* requires both domain-specific knowledge (e.g., runoff coefficients,

Model	Memorization				Understanding			Inference					Overall
	M-P	M-R	M-E	M-T	U-I	U-S	U-V	I-D	I-N	I-P	I-S	I-G	
Random	12.64	17.24	14.77	22.40	12.29	10.80	48.39	13.90	15.84	14.23	11.36	6.54	17.61
Expert (partial)	70.60	100.00	58.18	85.70	67.20	64.30	78.10	70.00	83.30	78.40	80.00	81.80	70.86
Mistral-7B-Instruct	24.44	21.31	36.10	37.12	21.96	29.25	48.84	34.99	21.28	25.90	45.18	21.18	29.10
Llama3-8B-Instruct	27.77	24.04	35.38	41.67	23.66	32.38	47.41	36.23	23.01	37.72	48.84	19.00	31.38
Baichuan2-7B-Chat	28.86	21.31	36.82	25.00	27.49	31.19	50.01	36.06	17.82	34.18	53.48	23.36	32.46
DeepSeek-LLM-7B-Chat	29.76	22.95	39.28	31.06	29.29	32.10	51.45	39.70	17.14	39.00	50.43	24.61	33.76
Phi-3.5-Mini-Instruct	30.08	28.96	47.65	50.76	25.97	37.01	46.12	37.63	25.79	47.42	47.13	23.05	34.22
Mistral-Nemo-Instruct	30.2	25.68	37.8	50.76	29.99	36.28	50.01	37.55	21.49	39.35	52.99	23.99	34.39
Baichuan2-13B-Chat	32.27	32.24	46.10	33.33	28.78	34.89	47.27	44.50	24.21	45.51	49.69	25.55	35.53
Marco-o1	34.77	33.33	35.91	36.36	33.66	36.43	49.74	43.51	27.62	35.46	66.06	28.04	37.32
Qwen2.5-3B-Instruct	37.14	36.61	43.07	54.55	33.33	39.60	50.65	42.51	37.58	39.70	61.66	29.28	39.67
Yi-1.5-9B-Chat	37.46	24.04	44.85	40.15	36.47	42.72	58.66	38.30	30.50	43.74	58.00	16.51	41.20
InternLM2.5-20b-Chat	39.80	40.44	42.23	39.39	35.98	42.50	53.90	41.85	43.55	44.02	64.22	29.60	42.26
Llama3-70B-Instruct	40.45	36.07	45.11	45.45	37.24	49.09	49.41	45.08	39.05	56.05	62.27	32.09	43.16
GLM-4-9b-Chat	41.85	43.72	51.48	50.00	37.14	48.69	51.56	46.48	22.69	59.24	58.24	26.17	43.72
InternLM2.5-7B-Chat	41.51	39.34	46.40	48.48	38.75	44.36	52.71	47.39	36.06	51.52	65.08	25.23	43.75
Qwen2.5-7B-Instruct	46.4	48.09	53.14	59.85	45.21	49.97	53.68	50.54	36.90	54.49	66.18	31.78	48.21
Qwen2.5-14B-Instruct	47.81	47.54	47.05	56.82	47.68	51.64	55.74	48.80	49.21	44.44	72.04	33.96	49.53
Yi-1.5-34b-Chat	50.24	45.36	51.44	66.67	50.57	52.67	62.28	46.82	36.32	53.86	66.91	30.22	51.83
Qwen2.5-32B-Instruct	55.32	53.55	52.05	65.91	54.31	57.80	61.53	55.17	55.19	55.34	73.50	50.16	56.35
DeepSeek-V3	56.39	54.10	57.05	59.09	52.06	58.61	59.66	**61.29**	**65.25**	61.15	**76.19**	**53.27**	57.43
Qwen2.5-72B-Instruct	60.15	56.83	**58.48**	**70.45**	60.45	61.91	62.69	55.09	52.78	59.66	73.02	51.09	60.32
Qwen2-72B-Instruct	**62.61**	**57.92**	58.11	66.67	**63.23**	**65.26**	**68.67**	56.16	45.65	**63.55**	73.63	51.09	**62.72**
GPT-3.5-Turbo	31.20	31.15	39.55	36.36	28.67	36.10	49.16	40.45	19.34	35.67	52.75	22.43	34.43
GLM-4-Flash	43.90	47.54	53.30	59.09	38.97	52.27	51.56	47.15	23.27	62.63	58.61	30.84	45.54
GPT-4o-mini	46.98	45.90	**59.89**	54.55	41.99	56.04	48.04	56.82	29.09	61.57	63.37	39.25	48.19
GPT-4o	47.38	44.26	47.16	45.45	47.09	51.81	59.71	55.33	42.14	53.29	**73.26**	42.99	50.01
GLM-4-Air	48.07	44.26	53.30	54.55	47.98	53.93	56.91	55.58	30.66	55.84	68.13	28.97	50.05
Claude-3.5-Sonnet	52.49	50.82	59.09	50.00	50.51	55.97	61.45	**59.80**	47.33	**69.43**	67.77	53.27	54.92
Gemini-2.0-Flash	52.30	57.38	55.68	54.55	52.95	52.72	62.21	55.58	**66.51**	60.51	71.06	52.34	55.33
Qwen-Turbo	54.85	50.82	51.36	59.09	55.04	55.06	64.46	55.33	46.54	53.08	71.06	39.25	55.76
Qwen-Plus	**63.83**	**60.66**	58.30	**68.18**	**63.78**	**63.82**	**67.21**	53.85	54.09	59.24	**73.26**	**56.07**	**63.21**

Table 3: Zero-shot performance of selected LLMs on cognitive tasks involving multiple-choice questions. Expert results are reported based on a subset of 1,500 samples. The best performance within each model series is highlighted in bold, and the second-best is underlined. Randomized results are based on seed 16.

crop water consumption) and multi-step quantitative reasoning. Such tasks reveal fundamental limitations in current LLM architectures, which struggle to integrate background knowledge with chained numerical computation, highlighting the need for structured reasoning, symbolic grounding, or tool-augmented approaches in complex, domain-specific scenarios.

Optimal LLM performance remains below expert level. AgriEval covers diverse subfields and includes tasks requiring factual recall, knowledge validation, and reasoning under complex agronomic conditions. To evaluate LLMs against human expertise, we construct an expert validation set by uniformly sampling 1,500 questions across all categories. Three agricultural experts with PhDs are recruited to annotate and answer the questions. As shown in Figure 1(*Right*), the experts achieve an average accuracy of 70.86%, outperforming the best-performing LLM by 5.06%. This gap highlights that, despite recent advances, LLMs still struggle with high-level reasoning and domain-specific knowledge in agricultural tasks.

Notably, expert performance is also imperfect. While each expert possesses deep knowledge in specific areas, accuracy declines on questions outside their core domains. This reveals a shared limitation for both humans and LLMs: difficulty in generalizing across the full breadth of agricultural knowledge.

Further Analysis

Larger models achieve better performance but exhibit diminishing returns. As shown in Table 3, we evaluate the performance of Qwen2.5 models across scales from 3B to 72B and observe larger models generally achieve higher accuracy, aligning with the scaling law (Kaplan et al. 2020). However, the performance gains exhibit diminishing returns as the model size increases beyond 14B. This suggests that simply scaling up parameters is insufficient for solving complex domain-specific tasks, especially when domain adaptation or reasoning capability becomes the bottleneck.

Instruction tuning significantly improves model performance and robustness. As shown in Table 3, instruction-tuned models consistently outperform their base counterparts, with an average accuracy gain of 10.60%. This improvement stems from supervised fine-tuning and alignment techniques that enhance instruction following and response quality. The performance gain is observed consistently across all question types, indicating stronger robustness in handling diverse task formats.

Cross-lingual gaps challenge model generalization. Chinese-oriented LLMs perform moderately well on AgriEval, while English-oriented models like Llama (Touvron et al. 2023) consistently underperform. This reveals challenges in cross-lingual generalization, as English-

Model	G-CA	G-QA	G-PS	Overall
Mistral-7B-Instruct	1.65	1.59	1.69	1.61
Llama3-8B-Instruct	1.97	1.86	1.95	1.88
Mistral-Nemo-Instruct	2.03	1.90	2.01	1.92
Phi-3.5-Mini-Instruct	2.20	2.05	2.31	2.09
Baichuan2-7B-Chat	2.35	2.22	2.37	2.25
Llama3-70B-Instruct	2.30	2.33	2.50	2.35
DeepSeek-7B-Chat	2.38	2.36	2.33	2.36
Baichuan2-13B-Chat	2.49	2.55	2.56	2.55
Yi-1.5-9B-Chat	2.73	2.66	2.74	2.67
Yi-1.5-34B-Chat	2.71	2.73	2.75	2.73
Qwen2.5-3B-Instruct	2.75	2.75	2.74	2.75
GLM-4-9b-Chat	2.82	2.80	2.81	2.81
Marco-o1	2.90	2.83	2.86	2.84
InternLM2.5-7B-Chat	3.03	2.92	2.99	2.93
Qwen2.5-7B-Instruct	3.04	3.06	3.06	3.06
Qwen2.5-14B-Instruct	3.05	3.07	3.04	3.07
InternLM2.5-20B-Chat	3.14	3.08	3.10	3.09
Qwen2.5-32B-Instruct	3.13	3.21	3.17	3.20
Qwen2.5-72B-Instruct	**3.30**	3.27	3.15	3.26
Qwen2-72B-Instruct	3.27	**3.32**	**3.22**	**3.30**
Claude-3.5-Sonnet	2.31	2.30	2.38	2.31
GPT-3.5-Turbo	2.38	2.29	2.38	2.31
GLM-4-Flash	2.74	2.78	2.85	2.79
GPT-4o-mini	2.97	2.96	2.99	2.97
GLM-4-Air	2.95	3.01	3.00	3.01
GPT-4o	3.02	3.16	3.16	3.16
DeepSeek-V3	3.17	3.20	3.23	3.20
Gemini-2.0-Flash	3.36	3.26	3.39	3.29
Qwen-Turbo	3.29	3.34	3.36	3.34
Qwen-Plus	**3.48**	**3.47**	**3.46**	**3.48**

Table 4: Zero-shot performance of selected LLMs on open-ended Q&A. Scores are rated by GPT-4o (scale from 1 to 4).

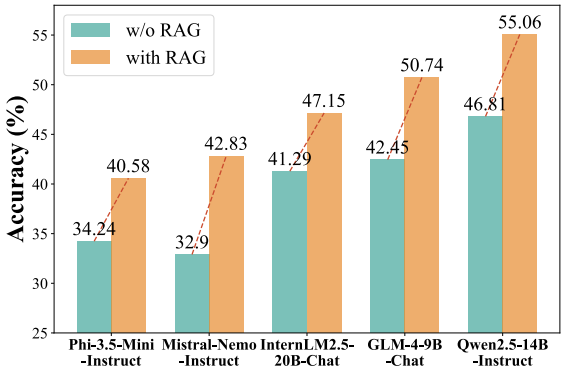

Figure 3: Effects of RAG.

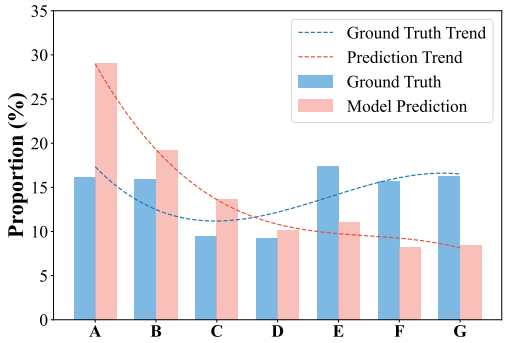

Figure 4: Exploratory analysis: effects of option order bias.

pretrained models struggle with Chinese domain-specific content such as crop terms, regulatory language, and regional expressions. These results highlight the need for targeted pretraining or fine-tuning to bridge language gaps in non-English, high-stakes domains like agriculture.

Exploration

CoT improves reasoning-intensive tasks but hinders performance on factual ones. To assess the effectiveness of CoT prompting, we compare model performance under zero-shot and CoT settings (Table 5). On average, CoT leads to a 3.51% performance drop, aligned with MMLU (Wang et al. 2024) and C-Eval (Huang et al. 2024). This suggests that CoT may introduce unnecessary reasoning steps when shallow pattern matching or factual recall is sufficient. However, a fine-grained analysis reveals that CoT brings significant benefits in specific scenarios. For numerical reasoning tasks, CoT improves accuracy by 9.81% on average, demonstrating its strength in guiding structured, multi-step computation. These improvements highlight CoT's potential in tasks that require step-wise reasoning or combinatorial decision-making. Taken together, these results suggest that the utility of CoT is highly task-dependent. Even though it may hinder performance on fact-based questions by introducing unnecessary complexity, it proves beneficial in inference-heavy contexts.

Few-shot learning cannot stably improve performance. We further explore the impact of in-context learning using a 5-shot setting, with results shown in Table 5. The results indicate that in-context learning yields inconsistent performance on AgriEval and does not always lead to improvements. We observe that model performance is highly sensitive to the relevance and quality of selected examples; semantically misaligned demonstrations may introduce noise and increase cognitive load. This suggests that in-context learning requires careful design in domain-specific tasks and that context effectiveness can be improved through semantically aligned example selection or demonstration filtering strategies.

RAG as an effective approach for rapid domain adaptation. To evaluate the impact of external knowledge, we construct a retrieval corpus from Chinese Wikipedia and randomly sample 1,000 examples across all categories to conduct RAG experiments. As shown in Figure 3, RAG consistently improves accuracy, with an average gain of approximately 4.0%. Notably, smaller models benefit the most, suggesting that external knowledge can partially compensate for limited model capacity. These results demonstrate the effectiveness of RAG in enhancing factual accuracy and knowledge grounding in agriculture domain tasks. However, varying performance gains across models highlight the need for better retrieval quality and more effective integration of retrieved information.

LLMs exhibit positional bias in multiple-choice tasks. To assess LLMs' sensitivity to answer order, we place mostly correct answers in later positions (e.g., D-G). In

Model	Prompt	Memorization				Understanding			Inference					Overall
		M-T	M-P	M-R	M-E	U-V	U-I	U-S	I-P	I-N	I-D	I-S	I-G	
Qwen2.5-3B -Instruct	Zero-Shot	37.14	36.61	43.07	54.55	33.33	39.60	50.65	42.51	37.58	39.70	61.66	29.28	39.67
	CoT	34.86↓	38.80↑	41.48↓	39.39↓	30.98↓	39.85↑	52.27↑	36.06↓	49.16↑	46.36↑	58.36↓	24.30↓	38.90↓
	Few-Shot	38.48↑	31.69↓	45.64↑	37.88↓	35.21↑	38.52↓	52.78↑	41.77↓	33.81↓	42.25↑	59.58↓	30.84↑	40.67↑
Qwen2.5-14B -Instruct	Zero-Shot	47.81	47.54	47.05	56.82	47.68	51.64	55.74	48.80	49.21	44.44	72.04	33.96	49.53
	CoT	45.31↓	40.44↓	44.77↓	41.67↓	44.37↓	48.92↓	57.84↑	45.33↓	64.99↑	47.13↑	64.84↓	38.01↑	48.39↓
	Few-Shot	51.57↑	42.62↓	46.89↓	53.03↓	50.13↑	53.17↑	57.25↑	50.87↑	51.68↑	48.20↑	69.72↓	40.50↑	52.07↑
GLM-4-9B -Chat	Zero-Shot	41.85	43.72	51.48	50.00	37.14	48.69	51.56	46.48	22.69	59.24	58.24	26.17	43.72
	CoT	39.48↓	45.36↑	47.65↓	44.70↓	35.80↓	48.09↓	52.92↑	43.09↓	56.29↑	56.97↓	60.68↑	23.36↓	43.78↑
	Few-Shot	38.25↓	42.62↓	42.12↓	53.03↑	35.03↓	44.64↓	49.89↓	40.53↓	22.17↓	48.05↓	53.97↓	24.61↓	40.07↓
GPT-3.5 -Turbo	Zero-Shot	31.20	31.15	39.55	36.36	28.67	36.10	49.16	40.45	19.34	35.67	52.75	22.43	34.43
	CoT	32.09↑	40.98↑	39.09↓	29.55↓	29.74↑	40.18↑	49.77↑	34.00↓	49.69↑	46.71↑	47.99↓	17.76↓	36.78↑
	Few-Shot	34.06↑	45.90↑	38.75↓	45.45↑	30.54↑	39.80↑	49.62↑	36.97↓	22.17↑	36.09↑	53.48↑	29.91↑	36.47↑
GPT-4o -mini	Zero-Shot	46.98	45.90	59.89	54.55	41.99	56.04	48.04	56.82	29.09	61.57	63.37	39.25	48.19
	CoT	40.91↓	40.98↓	47.61↓	40.91↓	37.33↓	48.72↓	54.41↑	30.77↓	41.19↑	47.56↓	50.92↓	32.71↓	43.29↓
	Few-Shot	48.36↑	52.46↑	62.27↑	72.73↑	42.92↑	57.78↑	52.47↑	58.31↑	26.89↓	54.99↓	64.10↑	36.45↓	49.63↑

Table 5: Comparison of five models on multi-choice cognitive tasks under zero-shot, few-shot, and CoT settings. ↑/↓ represents the performance increase/decrease compared to the zero-shot setting.

the biased setting, although 58.50% of correct answers appear later, models select them only 37.84% of the time, as shown in Figure 4. These results align with prior studies (Zheng et al. 2023; Du et al. 2024), which confirm that current LLMs favor positional heuristics over semantic reasoning. This calls for position-robust evaluation and training strategies, such as permutation augmentation and invariant prompting.

Error Analysis

In this section, we conduct an error analysis of GPT-4o-mini to uncover key limitations in domain-specific agricultural tasks and inform future improvements for LLM deployment. We sample 200 error cases and manually classify them into three categories: lack of knowledge, understanding error, and reasoning error. The distribution of error types is shown in Figure 5.

Lack of knowledge. The majority of errors are caused by missing domain-specific knowledge. In these cases, the model fails to answer correctly due to insufficient domain-specific knowledge, particularly in agronomy, aquaculture, and forestry. For instance, when asked about the optimal water temperature range for aquaculture species like groupers, the model fails to answer correctly due to missing information about species-specific thermal tolerances and breeding conditions. This highlights the need for stronger domain grounding and specialized pretraining.

Understanding errors. These account for 8% of cases and typically involve the model misinterpreting question intent or its own prior knowledge. For the former, models often fail to identify the "most relevant" option when all choices being contextually plausible, revealing limitations in comparative judgment. For the latter, models may initially provide accurate domain-specific explanations but introduce contradictions later in the response, indicating a lack of coherence during multi-step reasoning.

Reasoning errors. These are mostly found in numerical or procedural tasks involving biological quantities or resource planning. While CoT prompting improves step-by-step reasoning, the model still produces incorrect formulas or intermediate values.

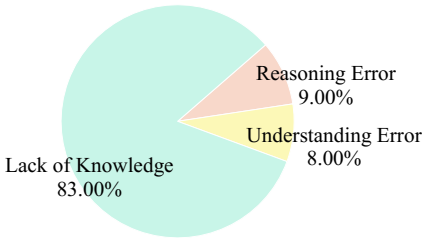

Figure 5: Overall error distribution for 200 annotated GPT-4o-mini errors.

Conclusion

As the largest benchmark designed for agricultural production, AgriEval spans most agronomy subfields, aligns with human professional-level testing formats, and provides a comprehensive cognitive classification. This enables a systematic evaluation of current models' capabilities relative to human experts in agriculture. Our evaluation of 51 commercial and open-source LLMs reveals that even top-performing models, such as Qwen-Plus, struggle with real-world production management. Through an in-depth analysis examining factors such as model size, version differences, language orientation, the effectiveness of few-shot and CoT prompting, the necessity of external knowledge retrieval, generation bias, cognitive ability levels, and common errors, we identify key performance drivers and suggest areas for improvement. We believe AgriEval will assist smart agriculture developers in addressing knowledge gaps in agricultural LLMs, enhancing model capabilities, and providing insights for constructing benchmarks in other specialized fields.

Acknowledgments

We thank the anonymous reviewers for their insightful comments. This work was supported in part by the Key Research and Development Program of Heilongjiang Province, China [2024ZX01A07], and the National Natural Science Foundation of China [72293584, 72431004].

References

Agathokleous, E.; Rillig, M. C.; Peñuelas, J.; and Yu, Z. 2024. One hundred important questions facing plant science derived using a large language model. *Trends in plant science*, 29(2): 210–218.

Bojar, O.; Buck, C.; Federmann, C.; Haddow, B.; Koehn, P.; Leveling, J.; Monz, C.; Pecina, P.; Post, M.; Saint-Amand, H.; et al. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, 12–58.

Chu, Z.; Chen, J.; Chen, Q.; Yu, W.; He, T.; Wang, H.; Peng, W.; Liu, M.; Qin, B.; and Liu, T. 2024. Navigate through Enigmatic Labyrinth A Survey of Chain of Thought Reasoning: Advances, Frontiers and Future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1173–1203. Bangkok, Thailand: Association for Computational Linguistics.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1): 37–46.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Du, L.; Sun, Z.; Ding, X.; Ma, Y.; Zhao, Y.; Qiu, K.; Liu, T.; and Qin, B. 2024. Causal-Guided Active Learning for Debiasing Large Language Models. *arXiv preprint arXiv:2408.12942*.

Gu, J.; Jiang, X.; Shi, Z.; Tan, H.; Zhai, X.; Xu, C.; Li, W.; Shen, Y.; Ma, S.; Liu, H.; et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

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*.

Huang, Y.; Bai, Y.; Zhu, Z.; Zhang, J.; Zhang, J.; Su, T.; Liu, J.; Lv, C.; Zhang, Y.; Fu, Y.; et al. 2024. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

Jiang, J.; Yan, L.; Liu, H.; Xia, Z.; Wang, H.; Yang, Y.; and Guan, Y. 2025. Knowledge assimilation: Implementing knowledge-guided agricultural large language model. *Knowledge-based systems*, 314: 113197.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kästing, M.; and Hänig, C. 2024. Assessing large language models in the agricultural sector: A comprehensive analysis utilizing a novel synthetic benchmark dataset. In *INFORMATIK 2024*, 1279–1286. Gesellschaft für Informatik eV.

Kpodo, J.; Kordjamshidi, P.; and Nejadhashemi, A. P. 2024. AgXQA: A benchmark for advanced Agricultural Extension question answering. *Computers and Electronics in Agriculture*, 225: 109349.

Kuska, M. T.; Wahabzada, M.; and Paulus, S. 2024. AI for crop production – Where can large language models (LLMs) provide substantial value? *Computers and Electronics in Agriculture*, 221: 108924.

Lee, H.; Phatale, S.; Mansoor, H.; Lu, K. R.; Mesnard, T.; Ferret, J.; Bishop, C.; Hall, E.; Carbune, V.; and Rastogi, A. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback.

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.

Li, H.; Chen, Y.; Ai, Q.; Wu, Y.; Zhang, R.; and Liu, Y. 2024a. Lexeval: A comprehensive chinese legal benchmark for evaluating large language models. *arXiv preprint arXiv:2409.20288*.

Li, H.; Zhang, Y.; Koto, F.; Yang, Y.; Zhao, H.; Gong, Y.; Duan, N.; and Baldwin, T. 2024b. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, 11260–11285. Bangkok, Thailand: Association for Computational Linguistics.

Li, J.; Xu, M.; Xiang, L.; Chen, D.; Zhuang, W.; Yin, X.; and Li, Z. 2024c. Foundation models in smart agriculture: Basics, opportunities, and challenges. *Computers and Electronics in Agriculture*, 222: 109032.

MacNish, T. R.; Danilevicz, M. F.; Bayer, P. E.; Bestry, M. S.; and Edwards, D. 2025. Application of machine learning and genomics for orphan crop improvement. *Nature communications*, 16(1): 982.

Navigli, R.; Conia, S.; and Ross, B. 2023. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2): 1–21.

Peng, R.; Liu, K.; Yang, P.; Yuan, Z.; and Li, S. 2023. Embedding-based retrieval with llm for effective agriculture information extracting from unstructured data. *arXiv preprint arXiv:2308.03107*.

Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 784–789.

Romera-Paredes, B.; and Torr, P. 2015. An embarrassingly simple approach to zero-shot learning. In *International conference on machine learning*, 2152–2161. PMLR.

Seaman, M. 2011. BLOOM'S TAXONOMY. *Curriculum & Teaching Dialogue*, 13.

Shahriar, S.; Corradini, M. G.; Sharif, S.; Moussa, M.; and Dara, R. 2025. The role of generative artificial intelligence in digital agri-food. *Journal of Agriculture and Food Research*, 20: 101787.

Silva, B.; Nunes, L.; Estevão, R.; Aski, V.; and Chandra, R. 2023. GPT-4 as an agronomist assistant? Answering agriculture exams using large language models. *arXiv preprint arXiv:2310.06225*.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tzachor, A.; Devare, M.; Richards, C.; Pypers, P.; Ghosh, A.; Koo, J.; Johal, S.; and King, B. 2023. Large language models and agricultural extension services. *Nature food*, 4(11): 941–948.

Wang, X.; Chen, G. H.; Song, D.; Zhang, Z.; Chen, Z.; Xiao, Q.; Jiang, F.; Li, J.; Wan, X.; Wang, B.; et al. 2023. Cmb: A comprehensive medical benchmark in chinese. *arXiv preprint arXiv:2308.08833*.

Wang, Y.; Ma, X.; Zhang, G.; Ni, Y.; Chandra, A.; Guo, S.; Ren, W.; Arulraj, A.; He, X.; Jiang, Z.; et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

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

Yang, B.; Zhang, Y.; Feng, L.; Chen, Y.; Zhang, J.; Xu, X.; Aierken, N.; Li, Y.; Chen, Y.; Yang, G.; et al. 2025. Agrigpt: A large language model ecosystem for agriculture. *arXiv preprint arXiv:2508.08632*.

Yang, X.; Gao, J.; Xue, W.; and Alexandersson, E. 2024. Pllama: An open-source large language model for plant science. *arXiv preprint arXiv:2401.01600*.

Yu, W.; Yang, Z.; Li, L.; Wang, J.; Lin, K.; Liu, Z.; Wang, X.; and Wang, L. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Zhang, G.; Du, X.; Chen, B.; Liang, Y.; Luo, T.; Zheng, T.; Zhu, K.; Cheng, Y.; Xu, C.; Guo, S.; et al. 2024. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2401.11944*.

Zheng, C.; Zhou, H.; Meng, F.; Zhou, J.; and Huang, M. 2023. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Zhou, Y.; and Ryo, M. 2024. Agribench: A hierarchical agriculture benchmark for multimodal large language models. In *European Conference on Computer Vision*, 207–223. Springer.