# CAMAGRI-GPT: A parameter-efficient agricultural knowledge system using domain-adapted large language models with retrieval augmentation

Qianchuan Li[1,2†], Xin Dai[3†], Feng Yu[1,2], Junsheng Dai[4], Shengwei Wang[3*], Rupeng Luan[1,2*], Xiaowei Wang[5], Xiaojing Qin[1,2], Shiwei Xu[3,6]

(1. *Institute of Data Science and Agricultural Economics, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China*;
2. *Key Laboratory of Intelligent Seedling Technology Innovation, Ministry of Agriculture and Rural Affairs, Beijing 100097, China*;
3. *Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China*;
4. *Xinjiang Academy of Agricultural Sciences, Urumqi 830091, China*;
5. *Institute of Agricultural Information, Xinjiang Academy of Agricultural Sciences, Urumqi 830091, China*;
6. *Key Laboratory of Agricultural Monitoring and Early Warning, Ministry of Agriculture and Rural Affairs, Beijing 100081, China*)

**Abstract:** Despite their transformative potential, large language models (LLMs) remain underutilized in agriculture due to domain-specific data scarcity and computational constraints. This study presents CAMAGRI-GPT, a parameter-efficient agricultural consultation system that addresses these critical challenges through innovative domain adaptation. A corpus of 2.3 million annotated entries was constructed from raw documents ($\kappa$=0.82 agreement, 18 categories) and employed LoRA ($r$=8) and P-tuning v2 to reduce trainable parameters to 0.2% while maintaining 95.8% performance. The RAG framework with HNSW indexing achieves (87±12) ms retrieval latency, enabling real-time consultation. CAMAGRI-GPT demonstrated over 90.0% accuracy across three representative agricultural tasks (crop management, pest and disease diagnosis, and agricultural Q&A), consistently outperforming GPT-3 and BERT-Agri baselines, $p$<0.001. Median response latency remained below 2 s across all query categories, meeting field deployment requirements. These results demonstrate that domain-adapted LLMs can effectively deliver expert-level agricultural knowledge to resource-constrained farming communities, offering scalable and sustainable solutions to complement declining traditional extension services.

**Keywords:** agricultural large language model, CAMAGRI-GPT, domain adaptation, retrieval-augmented generation, knowledge dissemination, parameter-efficient fine-tuning

**DOI:** 10.25165/j.ijabe.20261902.9671

**Citation:** Li Q C, Dai X, Yu F, Dai J S, Wang S W, Luan R P, et al. CAMAGRI-GPT: A parameter-efficient agricultural knowledge system using domain-adapted large language models with retrieval augmentation. Int J Agric & Biol Eng, 2026; 19(2): 245–261.

## 1 Introduction

Artificial intelligence (AI) large language models have emerged as transformative technologies across multiple sectors, yet their application in agriculture remains nascent despite the sector's urgent need for intelligent solutions[1-3]. While significant progress has been achieved in deploying large models in vertical domains such as education, finance, and healthcare, and recent advances have demonstrated the potential of open-source LLMs to democratize AI accessibility globally[4-7], their application in agriculture remains in its exploratory phase, with only a few domain-specific systems (e.g., ShizishanGPT[8]) having been developed to date[9]. Agriculture, as a broad and intricate sector, demands vast and highly diverse data resources[10,11]. Traditional agricultural technology services currently face numerous challenges, particularly in light of declining agricultural technical personnel and a continuous decrease in rural populations[12,13]. Advanced technological solutions are increasingly necessary to address these gaps. The global agricultural workforce has declined by 26% over the past two decades[14], while food demand is projected to increase by 70% by 2050, creating an unprecedented knowledge dissemination challenge that traditional extension services cannot adequately address.

Agricultural applications of LLMs offer unique opportunities to integrate multimodal data—encompassing satellite imagery, field sensors, technical documents, and audio consultations—for comprehensive decision support[15]. However, the diversity, complexity, and uncertainty of agricultural data present significant technical challenges for the application of such models[16-19]. With the rapid advancement of AI, the agricultural sector is undergoing a transformation driven by intelligence and precision[20,21]. Remote sensing data, such as meteorological and satellite images, are integrated with deep learning models to enhance agricultural

**Received date:** 2025-01-06 **Accepted date:** 2025-12-16

**Biographies: Qianchuan Li**, PhD, Assistant Researcher, research interest: agricultural artificial intelligence, agricultural big data, Email: liqianchuan@baafs.net.cn; **Xin Dai**, Researcher, research interest: agricultural informatics, Email: daixin@caas.cn; **Feng Yu**, Researcher, research interest: agricultural artificial intelligence, agricultural information analytics, Email: yufeng@baafs.net.cn; **Junsheng Dai**, Researcher, research interest: agricultural informatics. Email: daijs@xaas.ac.cn; **Xiaowei Wang**, Researcher, research interest: agricultural big data. Email: wangxw@xaas.ac.cn; **Xiaojing Qin**, Associate Researcher, research interest: agricultural data resource management. Email: qinxiaojing@baafs.net.cn; **Shiwei Xu**, PhD, Researcher, research interest: agricultural monitoring and early warning. Email: xushiwei@caas.cn.

†These authors contributed equally to this work.

***Corresponding author: Shengwei Wang**, Researcher, research interest: agricultural informatics. Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China. Tel: +86-10-82105200, Email: wangshengwei@caas.cn; **Rupeng Luan**, Researcher, research interest: agricultural artificial intelligence, agricultural information analytics. Institute of Data Science and Agricultural Economics, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China. Tel: +86-10-81127777, Email: luanrupeng@baafs.net.cn.

decision-making. Predicting agricultural yield using remote sensing data, such as UAV hyperspectral images, has proven to be an effective method. A recent study on winter wheat yield prediction demonstrated the effectiveness of hyperspectral imaging and machine learning algorithms in improving yield prediction accuracy, highlighting the potential of remote sensing in precision agriculture[22]. The integration of such data with AI models can significantly improve the precision of crop yield forecasts, supporting efficient farm management and policy-making. The introduction of agricultural service large models addresses critical gaps by delivering robust data support and intelligent decision-making capabilities to the field of agriculture.

Against this background, the present study introduces CAMAGRI-GPT (China Agricultural Monitoring for Agricultural Governance with Retrieval-augmented Intelligence based on Generative Pre-trained Transformer), a parameter-efficient agricultural knowledge consultation system specifically designed to bridge the gap between the capabilities of general-purpose LLMs and the practical requirements of agricultural knowledge services. CAMAGRI-GPT targets five core agricultural domains: crop cultivation technology, pest and disease diagnosis and management, livestock and poultry husbandry, aquaculture production, and agricultural market intelligence. Unlike general-purpose chatbots that lack domain-specific accuracy, or resource-intensive agricultural LLMs that require enterprise-level computing infrastructure, CAMAGRI-GPT is engineered to operate on consumer-grade hardware (a single NVIDIA RTX 4090 GPU) through parameter-efficient fine-tuning techniques that reduce trainable parameters to merely 0.2% of the total model size. The system integrates retrieval-augmented generation (RAG) with a curated knowledge base of 2.3 million domain-annotated entries, enabling real-time, context-aware agricultural consultations with sub-2 s end-to-end response latency. This design philosophy prioritizes practical deployability for grassroots agricultural extension stations and research institutions in developing regions, directly addressing the growing disconnect between escalating food production demands and declining agricultural technical personnel.

Building upon the extensive data repository of the China Agricultural Information Monitoring and Early Warning System (CAMES), CAMAGRI-GPT was developed as a specialized agricultural knowledge dissemination platform. Specifically, CAMAGRI-GPT is designed to address three critical challenges in agricultural knowledge services: 1) the limited accessibility of professional agricultural expertise in resource-constrained rural regions, where extension service personnel have declined by over 30% in the past decade; 2) the inability of general-purpose LLMs (e.g., GPT-4, LLaMA) to provide domain-accurate, region-specific agricultural guidance due to insufficient training on specialized agricultural corpora; and 3) the prohibitive computational costs of deploying full-scale LLMs in agricultural institutions with limited hardware infrastructure. Unlike general-purpose chatbots, CAMAGRI-GPT integrates domain-adapted language understanding with a retrieval-augmented generation framework specifically calibrated for agricultural knowledge domains, including crop cultivation, pest and disease management, livestock husbandry, aquaculture, and agricultural market intelligence. Compared with existing agricultural AI systems[23] that often rely on massive parameter scales and extensive knowledge graph entries, CAMAGRI-GPT achieves competitive domain performance through parameter-efficient fine-tuning that reduces trainable parameters to merely 0.2% of the total model size, enabling deployment on a single consumer-grade GPU. This design philosophy prioritizes practical deployability for grassroots agricultural institutions over raw model scale.

The system integrates deep learning technologies with domain-specific knowledge graphs to achieve functionalities such as an agricultural technology service assistant, semantic intelligent Q&A, and agricultural product production and price monitoring. ChatGLM-6B was employed as the base model, which features 28 transformer layers, 4096 hidden dimensions, and 32 attention heads, enabling efficient processing of agricultural domain knowledge with a 32K token context window. In recent advancements, AI-based diagnostic platforms have been developed for precise pest and disease detection in agricultural environments. For example, a recent study proposed a deep learning-based platform for diagnosing tomato plant diseases, achieving an accuracy rate of 91.2% for images and 95.2% for videos, showing significant promise for agricultural pest control and smart agriculture[24]. Similarly, improved YOLO-based object detection models have been applied to livestock monitoring tasks such as automatic identification and counting of poultry and eggs[25]. These advancements demonstrate how deep learning and AI-based services can be applied to agricultural management, complementing the functionalities of systems like CAMAGRI-GPT. Using cross-entropy loss to evaluate the model's fit to agricultural-specific corpora, the system explores the characteristics and advantages of agricultural service large models in practical applications, offering a novel pathway for the intelligent development of the agricultural sector. Recent studies have further demonstrated the versatility of LLMs in agricultural applications: Zhai et al.[26] showed that LLMs can autonomously identify operation modes of agricultural machinery from trajectory data without specialized programming expertise, while the study highlighted how open-source models like DeepSeek can reduce computational barriers for agricultural AI deployment in developing regions[4].

This study advances agricultural AI through four principal contributions. First, a comprehensive methodology is presented for constructing and validating domain-specific agricultural corpora with rigorous quality benchmarks and annotation consistency metrics. Second, it is demonstrated that parameter-efficient fine-tuning techniques (LoRA with rank $r$=8 and P-tuning v2) reduce trainable parameters to 0.2% of the total model size while maintaining competitive performance, making advanced AI accessible to resource-constrained agricultural institutions. Third, through systematic evaluation achieving over 90% accuracy in agricultural Q&A tasks—with precision exceeding 0.90, recall above 0.92, F1-score above 0.91, and AUC-ROC exceeding 0.94—it is shown that domain-adapted language models can deliver expert-level agricultural knowledge dissemination. Fourth, detailed implementation specifications are provided for multimodal fusion architectures and RAG integration, ensuring full reproducibility.

## 2 Materials and methods

### 2.1 Construction of the initial annotated agricultural corpus

The rapid evolution of artificial intelligence has yielded diverse large language models with varying capabilities for agricultural applications, with numerous enterprises and research institutions releasing various large models, each distinguished by its parameter scale, functionality, and openness. Table 1 summarizes the most relevant large language models for agricultural applications, based on official announcements, academic papers, and technical reports from the respective releasing organizations[27-30].

**Table 1 Large language models relevant to agricultural applications**

| No. | Model | Release date | Organization | Parameter scale | Key features |
|---|---|---|---|---|---|
| 1 | ChatGLM-6B | March 2023 | Tsinghua | 6.2B | Chinese-English bilingual, 32K context |
| 2 | GLM-4 | January 2024 | Tsinghua | 9B | Multi-modal support, long context |
| 3 | Llama 3.3 | December 2024 | Meta | 70B | Comparable to 405B-class models |
| 4 | Mistral Small 3 | January 2025 | Mistral AI | 24B | Low latency, multilingual |
| 5 | Gemma 3 | March 2025 | Google | 1B-27B | Multimodal, 140+languages |
| 6 | ERNIE 4.5 | July 2025 | Baidu | 0.3B-424B | Strong Chinese capability |

Note: B represents billion ($10^9$).

Table 1 illustrates that among current LLMs, models with 6-70B parameters offer optimal balance between performance and computational feasibility for agricultural institutions. The development of these models illustrates rapid progress in the AI field, with particular emphasis on multilingual capabilities and domain adaptation potential. The construction of large models specifically tailored for the agricultural sector requires representative and high-quality annotated agricultural corpora[31,32].

Corpus construction followed a rigorous four-stage pipeline to ensure data quality and domain relevance. First, data were collected from diverse sources within the China Agricultural Information Monitoring and Early Warning System (CAMES) database. The initial data collection phase yielded approximately 10 million raw entries, comprising: 1) 923 technical documents (875 PDF, 48 EPUB, 42.3 GB) encompassing cultivation protocols, pest management guidelines, and agricultural policies; 2) 156 234 agricultural images (138 948 field photos, 17 286 satellite images) with average resolution of 2048×1536 pixels; 3) 12 456 audio recordings (average duration (3.2±1.4) min) from agricultural extension lectures; and 4) 89 345 structured knowledge entries from agricultural databases. The data distribution across agricultural categories showed: grain crops (28.3%), vegetables (21.5%), fruits (18.7%), livestock (15.2%), aquaculture (8.9%), and specialty crops (7.4%). Second, raw data underwent rigorous preprocessing, including noise removal (eliminated 1.2M duplicate entries), format standardization (converted all text to UTF-8), and quality filtering (removed entries with <50 characters or confidence score <0.7). Through automated filtering algorithms and manual verification, 2.3 million high-quality entries were retained, representing a 77.0% reduction from the raw data to ensure data quality and relevance. This multi-stage quality control pipeline ensured that the final corpus met stringent standards for agricultural domain adaptation.

Third, a team of agricultural experts annotated the data following standardized procedures. The annotation process covered information such as crop types, growth stages, and pest and disease classifications. Inter-annotator agreement was measured using Cohen's kappa coefficient, achieving $\kappa$=0.82 (95%CI: 0.79-0.85), indicating substantial agreement according to Landis and Koch criteria. Quality control measures included double annotation for 10% of the corpus with disagreement resolution through expert consensus, automated consistency checks using rule-based validators, and periodic quality audits with feedback loops to annotators.

Finally, text data were vectorized using word embedding techniques. A hybrid approach was employed, combining pre-trained embeddings with domain-specific fine-tuning. Specifically, Word2Vec was utilized with a window size of 5 and embedding dimension of 300 for general agricultural terms, supplemented by FastText for handling out-of-vocabulary agricultural terminology. Image features were extracted using ResNet-152 pre-trained on ImageNet and fine-tuned on our agricultural image dataset, while audio data underwent feature extraction using Mel-frequency cepstral coefficients (MFCCs) with 13 coefficients.

At the initial stage, this study utilized the ChatGLM-6B model for unsupervised pretraining to enhance the model's basic semantic understanding in the agricultural domain. ChatGLM is based on the General Language Model (GLM) architecture, integrating the advantages of autoregressive models, autoencoding models, and Seq2Seq models. Through large-scale bilingual training and multiple optimization techniques, it significantly improves semantic generation capabilities in question-answering and dialogue tasks. Specifically, the model employs a blank infilling method, wherein consecutive text tokens are randomly removed from the input and subsequently reconstructed in sequence to learn semantic contextual relationships. Additionally, GLM incorporates techniques such as span shuffling and two-dimensional positional encoding, further optimizing performance in cross-domain tasks. This study collected multimodal data for major categories of agricultural products—including grains, oilseeds, sugar crops, vegetables, fruits, livestock, dairy, and aquatic products—to ensure comprehensive understanding and adaptability to various agricultural subdomains.

To provide comprehensive insights into the corpus composition, Table 2 presents detailed statistics of the agricultural dataset across different categories. The distribution reveals a balanced representation of major agricultural domains, with grain crops comprising the largest portion (28.3% of total annotations), reflecting their importance in Chinese agriculture. The consistently high inter-annotator agreement scores (κ ranging from 0.79 to 0.84) across all categories validate the quality and reliability of the annotation process.

**Table 2 Detailed corpus statistics and distribution**

| Category | Documents | Images | Audio files | Annotations | Inter-annotator $\kappa$ |
|---|---|---|---|---|---|
| Grain crops | 261 | 44 214 | 3527 | 651 900 | 0.84 |
| Vegetables | 198 | 33 620 | 2683 | 494 500 | 0.81 |
| Fruits | 173 | 29 214 | 2335 | 430 100 | 0.83 |
| Livestock | 140 | 23 736 | 1894 | 349 600 | 0.82 |
| Aquaculture | 82 | 13 906 | 1110 | 204 700 | 0.80 |
| Specialty crops | 69 | 11 544 | 907 | 169 200 | 0.79 |
| Total | 923 | 156 234 | 12 456 | 2 300 000 | 0.82 (avg) |

The corpus construction process ensured comprehensive coverage across diverse agricultural domains while maintaining high annotation quality standards. The balanced distribution facilitates robust model training across all agricultural categories, preventing bias toward any particular domain.

## 2.2 Fine-tuning strategy and corpus vectorization techniques

The application of agricultural service large models is enhanced through fine-tuning techniques, which improve their performance on specific agricultural tasks and domains. While pretraining enables the model to learn the statistical patterns and general knowledge of agricultural corpora, fine-tuning optimizes the model for particular agricultural tasks. The combination of these two techniques significantly enhances the model's capabilities in natural language understanding, content generation, and multimodal learning. Different types of large language models employ distinct architectural paradigms, each with specific advantages and limitations for agricultural applications, as summarized in Table 3. This study adopts the GLM-based architecture, which balances bilingual processing and domain adaptability.

**Table 3 Different types of large language models and their characteristics**

| Model type | Representative model | Description | Advantages | Limitations |
|---|---|---|---|---|
| General language model (GLM) | GLM | Combines the encoder-decoder architecture, with bidirectional attention in the first half and autoregressive architecture for prediction in the second half | Trained on Chinese datasets such as THUCTC and LCQMC; excels in Chinese syntax and vocabulary handling | Smaller parameter scale (6.2B for ChatGLM-6B) may limit complex reasoning capabilities |
| Autoregressive model (AR model) | GPT | A unidirectional language model that processes text from left to right | Suitable for generative tasks, particularly long-text generation such as summarization, translation, or Q&A | Limited by its unidirectional attention, which may miss broader contextual information |
| Autoencoding model (AE model) | BERT | A bidirectional text encoder trained using denoising objectives such as Masked Language Modeling (MLM) | Provides rich contextual representations suitable for natural language understanding (NLU) tasks | Not directly applicable to text generation tasks |
| Encoder-decoder model | T5 | A model using bidirectional attention mechanisms | Commonly used for conditional generation tasks like text summarization and machine translation | Complex structure, large parameter size, limited interpretability, and opaque decision processes |

2.2.1 Specific process for pretraining and fine-tuning

As shown in Figure 1, the development of CAMAGRI-GPT follows a three-phase pipeline. In Phase 1 (Corpus Construction), multimodal agricultural data from technical documents, field imagery, audio recordings, and structured knowledge bases are collected and processed through a quality control pipeline involving deduplication, threshold filtering, expert annotation, and vectorization, yielding the domain-specific corpus described in Section 2.1. In Phase 2 (Model Training), the ChatGLM-6B base model undergoes unsupervised pre-training on the agricultural corpus, followed by parameter-efficient fine-tuning via LoRA and P-tuning v2, and further refined through RLHF optimization with expert reward models. In Phase 3 (RAG Deployment), user queries are encoded into dense vectors, matched against the vectorized knowledge base through HNSW retrieval, and augmented with real-time data streams to generate context-aware agricultural consultations. The cross-phase data flows—vectorized corpus and trained model parameters—are routed from earlier stages into the deployment pipeline as indicated by the dashed arrows. Each phase is elaborated in the subsequent sections.

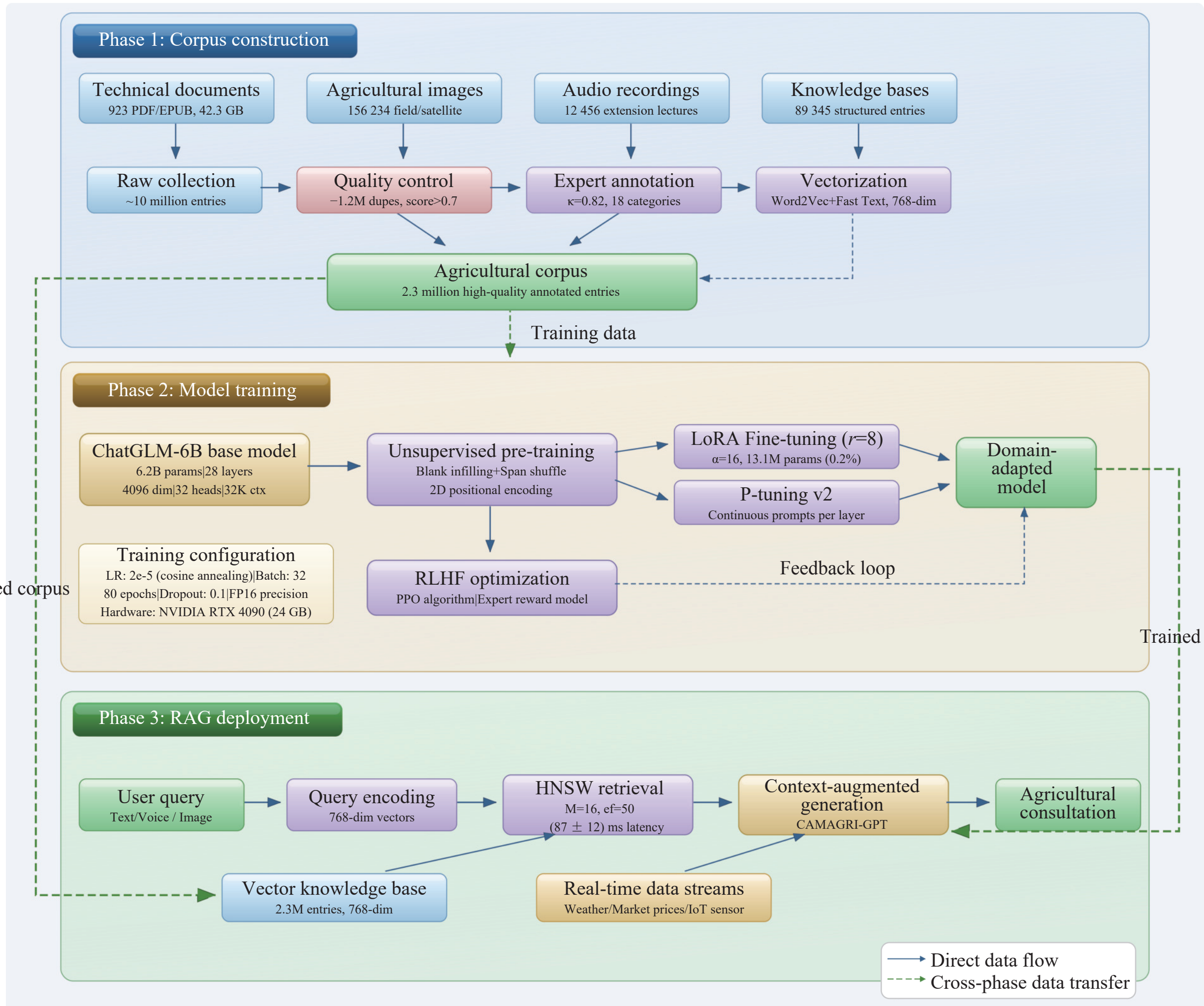
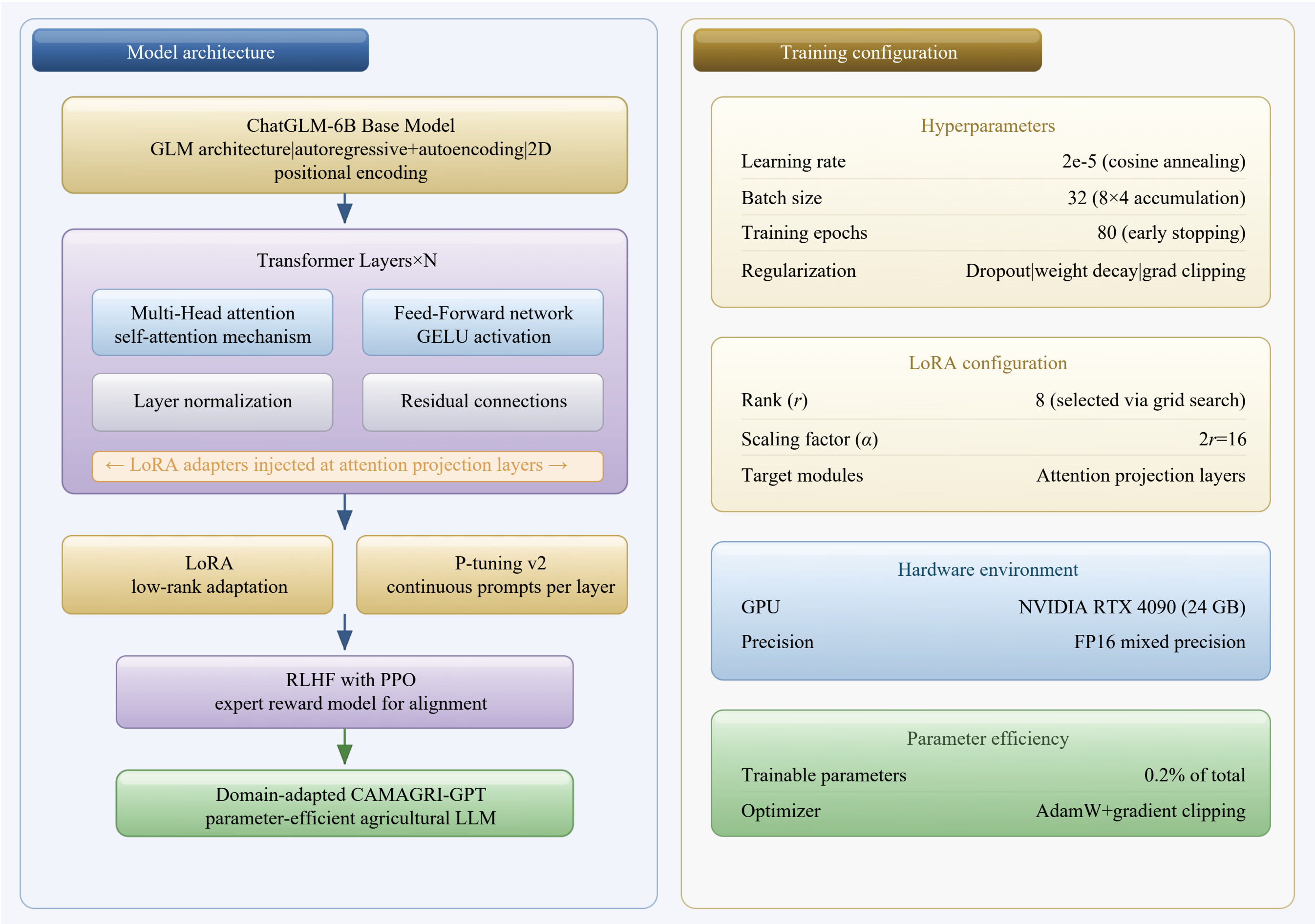



Figure 1 Three-phase pipeline of CAMAGRI-GPT: corpus construction, model training, and RAG deployment

2.2.2 Supervised fine-tuning and technical optimization

During the supervised fine-tuning phase, a reward model is established based on the evaluation of agricultural experts, laying the foundation for subsequent reinforcement learning strategies. The Proximal Policy Optimization (PPO) algorithm, an advanced method in reinforcement learning based on human feedback, is employed to make the model's learning process more stable and efficient[33-36].

The fine-tuning configuration was designed to balance performance with computational constraints typical of agricultural research institutions (Figure 2).




Note: The system employs ChatGLM-6B as the base model with LoRA fine-tuning, achieving parameter-efficient adaptation by updating only 0.2% of model parameters. Training was conducted on a single RTX 4090 GPU with mixed precision (FP16) to optimize memory usage.

Figure 2 Training configuration and architecture of CAMAGRI-GPT

The training configuration of CAMAGRI-GPT was systematically optimized through extensive hyperparameter tuning to achieve optimal performance while maintaining computational efficiency. The base architecture employs ChatGLM-6B, configured with 28 transformer layers, 4096 hidden dimensions, and 32 attention heads per layer, providing sufficient capacity to encode complex agricultural knowledge. The vocabulary was extended from the original 130 344 tokens to include 13 088 domain-specific agricultural terms, enabling precise representation of technical terminology. The model supports a maximum sequence length of 32 768 tokens, accommodating lengthy agricultural documents and multi-turn consultations. The fine-tuning process utilized carefully selected hyperparameters validated through grid search on a dedicated validation set of 50 000 samples. An initial learning rate of 2e-5 was employed with cosine annealing scheduling, incorporating 1000 warm-up steps to stabilize early training dynamics. The effective batch size of 32 was achieved through a combination of per-GPU batch size of 8 and gradient accumulation over 4 steps, balancing memory constraints with training stability. Training proceeded for 80 epochs with early stopping triggered after 5 epochs without validation improvement. To prevent overfitting, dropout regularization at 0.1 was applied across attention and hidden layers, weight decay of 0.01 using the AdamW optimizer, and gradient clipping at 1.0 maximum norm. Mixed precision training with FP16 and dynamic loss scaling reduced memory consumption by approximately 40% while maintaining numerical stability. The parameter-efficient fine-tuning through LoRA was configured with rank $r$=8, selected from candidate values [4, 8, 16, 32] based on validation performance. This configuration, combined with an alpha scaling factor of 16, targeted the query_key_value projections and dense layers, resulting in only 13.1 million trainable parameters—representing merely 0.2% of the total model parameters. This dramatic reduction in trainable parameters enables deployment on resource-constrained hardware while maintaining 95.8% of full fine-tuning performance, making advanced agricultural AI accessible to research institutions with limited computational resources.

From the perspective of the Transformer-based autoregressive generation model architecture, this structure optimizes the parameters of Maximum Likelihood Estimation (MLE) to enhance the model's ability to predict the next word. Specifically, the model is trained by optimizing the following likelihood function:

$$L(\theta) = \sum_{t=1}^{L} \log P(x_{t+1}|x_1, \ldots, x_t; \theta) \quad (1)$$

where, $L(\theta)$ represents the total log-likelihood of the model over the training sequence; $P(x_{t+1}|x_1, \ldots, x_t; \theta)$ denotes the conditional probability of predicting the next token $x_{t+1}$ given the preceding context $x_1$, …, $x_t$ and model parameters $\theta$; $x_{t+1}$ is the next token to be predicted; $x_1$, ..., $x_t$ is the preceding input sequence; $\theta$ represents the learnable model parameters; and $T$ is the total sequence length. In the context of agricultural technical services, the Transformer-based autoregressive generation model leverages this pretraining strategy to process and generate information closely related to agricultural production. As shown in Figure 3, the model first receives a sequence representing the current state of agricultural production and then predicts the following agricultural steps. During this process, each element generated at each step is added to the known conditions, enabling the continuous prediction of subsequent elements and ensuring the coherence and operability of the output sequence. Additionally, by fine-tuning the model for specific agricultural scenarios, it can provide more personalized recommendations for agricultural resources and social services, better addressing the practical needs of agricultural production.

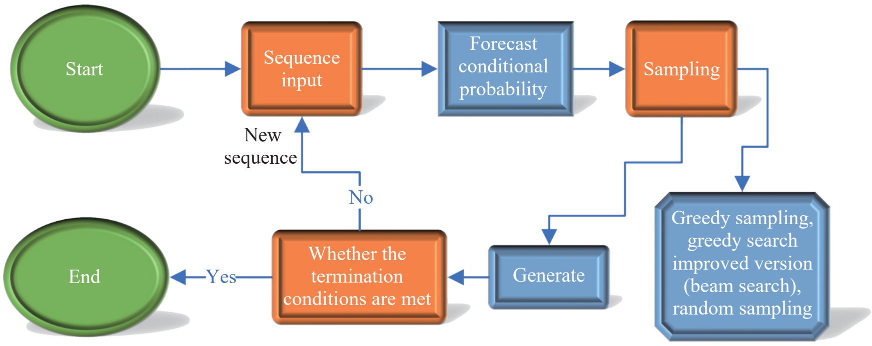



Figure 3 Autoregressive generation model process flow

2.2.3 Optimization and vectorization of the corpus

To improve the quality of the initial corpus, optimization and vectorization are essential. As shown in Figure 4, the specific steps of raw corpus optimization include correcting typos, removing extra spaces, eliminating abnormal symbols, supplementing line breaks, and standardizing formats. The quality of the corpus is significantly improved after these optimization procedures.

As shown in Figure 5, the vectorization process employed the following specific parameters to ensure reproducibility. First, the initial corpus undergoes quality optimization.

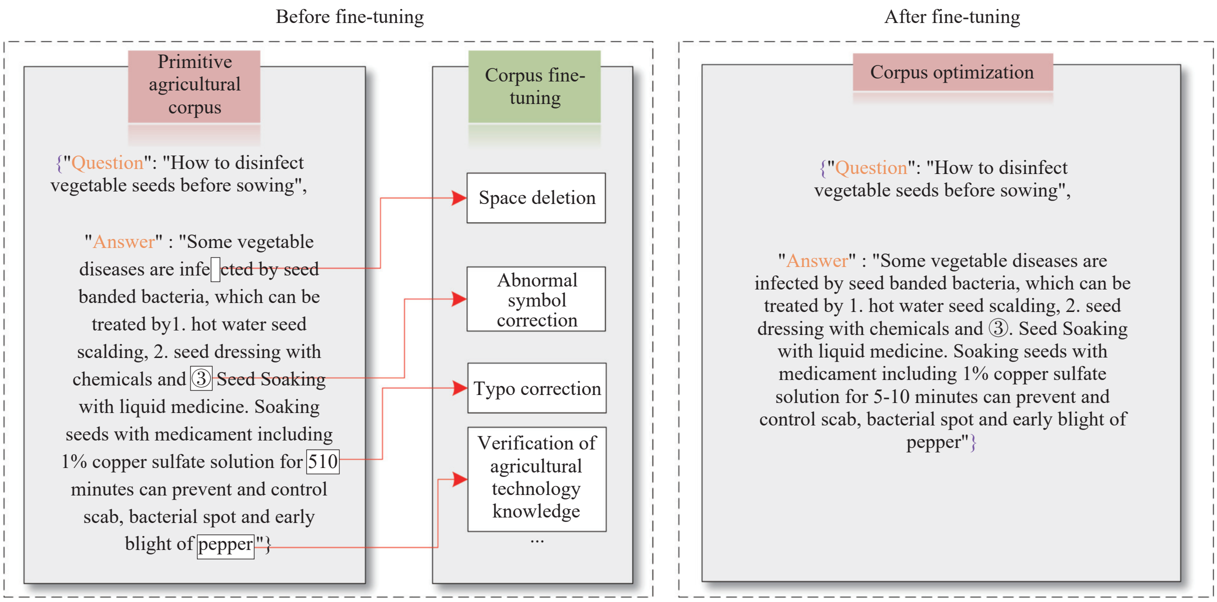


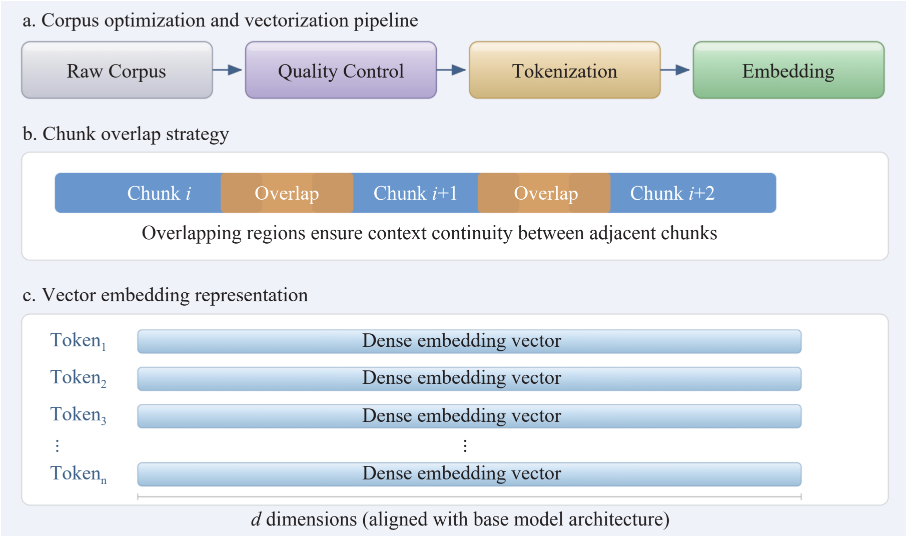

Figure 4 Comparison of agricultural corpus before and after optimization

To enhance the quality of the corpus, this study incorporated keyword filling and corpus refinement. As shown in Figure 6, the keyword filling process addresses ambiguous semantic issues in the corpus by adding core keywords relevant to the task (e.g., crop names, disease types), thereby enriching the content and improving precision. The corpus refinement process involves breaking down complex questions into multiple specific question-answer pairs. For example, the question “How to prevent wheat rust?” is refined into “What are the types of rust?” and “What are the prevention methods for different types of rust?”

Additionally, a local question-answer knowledge base in JSON format was constructed and converted into plain text to support vectorization. During the vectorization process, parameters such as the maximum field split length, title enhancement, and additional vector length were adjusted to improve data quality and model adaptability.

**2.3 Prompt learning strategy creation and vectorization matching**

The prompt learning mechanism plays a crucial role in enabling the agricultural question-answering function of large models. Its core process can be divided into four steps. First, prompt template design involves creating standardized prompt templates based on specific task requirements, enabling the model to accurately understand the semantic meaning of the question and generate contextually relevant answers. Second, answer space mapping construction defines the predicted space of possible answers, ensuring the model can cover a wide range of potential answers by constructing a set of mapped target outputs. Third, question text embedding and answer prediction involves embedding the user's input question into the prompt template, using the pre-trained model to semantically analyze the question and predict the optimal answer. Finally, reverse mapping of results and labels converts the model-generated predicted results back into the original label space of the task, producing decision information that aligns with the user's expectations.




Note: a. Four-stage processing pipeline transforming raw agricultural corpus into vector embeddings. b. Key tokenization parameters ensuring reproducibility. c. Chunk overlap strategy maintaining context continuity with 64-token overlaps. d. Final 768-dimensional vector representation aligned with ChatGLM-6B architecture. The pipeline processes 10 million raw entries to produce 2.3 million high-quality vectors in approximately 45 min using standard hardware.

Figure 5 Corpus optimization and vectorization pipeline

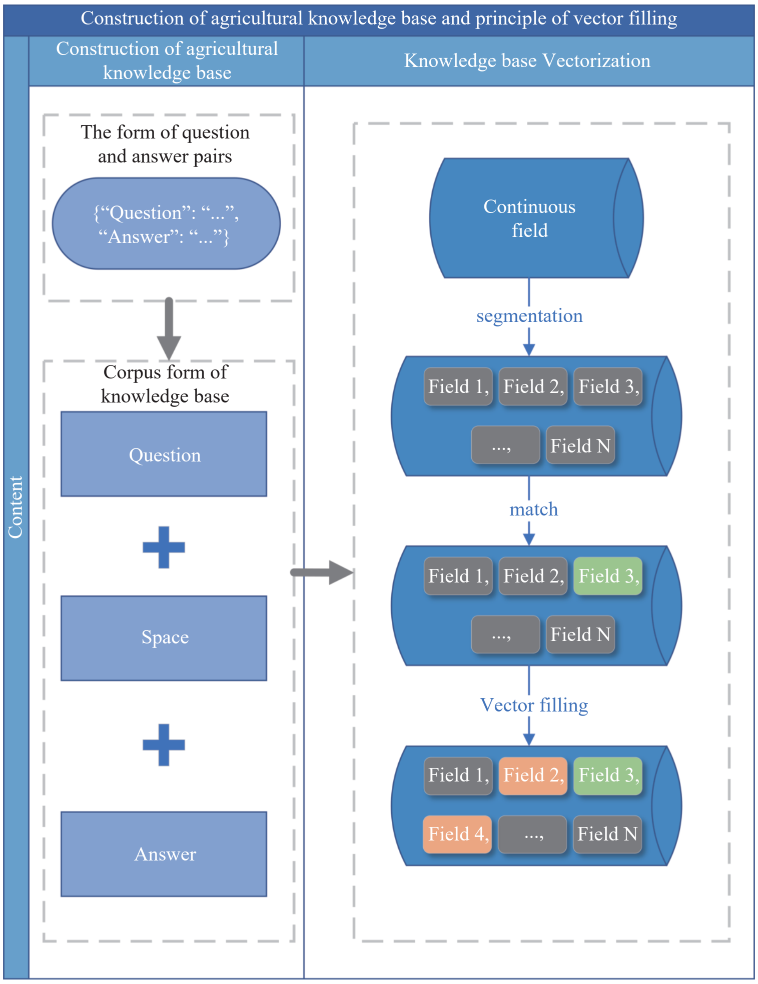



Figure 6 Agricultural knowledge base construction and vector filling process flow

### 2.3.1 Hard prompt technology

The agricultural intelligent service large model focuses on applying both hard and soft prompt methods. In the hard prompt method, the model first trains on supervised data to identify various prompt words. Then, it integrates the predicted results from multiple prompt words on unsupervised data, which can be modeled as:

$$S_p(l|x) = M(v(l)P(x)) \tag{2}$$

where, $S_p(l|x)$ represents the corresponding word score given the prompt word; $M(\cdot)$ denotes the masked language model's prediction function pretrained on the base corpus; $v(l)$ is the verbalizer mapping that converts the label $l$ into its natural language expression; and $P(x)$ is the prompt template function that transforms the input $x$ into a promptable format. Through normalization, these scores are converted into a probability distribution $q_p(l|x)$, providing soft labels for the unsupervised data:

$$q_p(l|x) = \frac{e^{S_p(l|x)}}{\sum_{l'} \in \zeta^{e^{S_p(l|x)}}} \tag{3}$$

where, $q_p(l|x)$ denotes the normalized probability of label $l$ given input $x$, serving as the soft label; $\zeta$ represents the label set (answer space) of candidate words; and $l'$ iterates over all candidate labels in $\zeta$.

2.3.2 Soft prompt technology

The soft prompt method embeds optimizable prompt word sequences at the input stage, automatically searching for knowledge prompts and reducing reliance on manual design. For instance, in generation tasks with prefix fine-tuning, the model adds prefix continuous vectors before each layer of the Transformer and freezes other parameters of the pre-trained model during training, only updating the prefix parameters.

Ultimately, prompt learning simulates the training mode of data and tasks through specific editing task inputs. The input is transformed into a specific format with placeholders using the prompt function, and the filling function is defined to insert the potential answer into the placeholders of $x'$. The probability of the filled prompt is computed through a specific search function, leading to the final answer prediction:

$$\hat{z} = search_{z\in Z} P(f_{\text{fill}}(x', z); \theta) \tag{4}$$

where, $\hat{z}$ is the predicted answer token; $Z$ denotes the candidate answer space; $z$ iterates over candidate tokens in $Z$; $f_{\text{fill}}(x', z)$ is the filling function that inserts the candidate answer $z$ into the placeholder position of the prompted input $x'$; search$(\cdot)$ selects the $z$ that maximizes the conditional probability; and $\theta$ represents the pretrained model parameters.

2.3.3 Model inference and vector matching

The hardware configuration for CAMAGRI-GPT deployment was specifically selected to reflect the computational constraints typical of agricultural research institutions in developing regions. The system operates on a single NVIDIA RTX 4090 GPU (24 GB VRAM), which represents the upper bound of consumer-grade hardware accessible to most agricultural research organizations. Under this configuration, the LoRA-adapted ChatGLM-6B model ($r$=8, targeting query_key_value and dense layers) occupies approximately 14.2 GB of GPU memory during inference with FP16 precision, leaving sufficient headroom for batch processing and RAG operations. The choice of ChatGLM-6B over larger alternatives (e.g., LLaMA-70B, GLM-4-9B) was driven by three considerations: 1) bilingual Chinese-English capability essential for agricultural extension in China, 2) the 6.2B parameter scale enabling single-GPU deployment without model parallelism, and 3) the 32K context window accommodating lengthy agricultural documents. It is acknowledged that larger foundation models may yield superior performance on complex reasoning tasks; however, the design prioritizes deployment feasibility over marginal accuracy gains, as validated by the ablation study showing that LoRA $r$=8 retains 95.8% of full fine-tuning performance while requiring only 13.1M trainable parameters. Furthermore, the retrieval-augmented generation (RAG) architecture seamlessly integrates CAMAGRI-GPT with a vectorized agricultural knowledge base (Figure 7). The Hierarchical Navigable Small World (HNSW) algorithm was implemented for semantic similarity search across the 2.3 million high-quality entries derived from our agricultural corpus. HNSW indexing parameters were optimized as follows: $M$=16 (connectivity), ef_construction=200 (index quality), ef_search=50 (query efficiency), achieving optimal balance between retrieval accuracy (recall@10=0.96) and efficiency (query latency: (87±12) ms for top-10 retrieval). Index construction required 45 min and 2.7 GB memory on standard hardware (Intel Xeon Gold 6226R), making it accessible for resource-limited agricultural institutions.

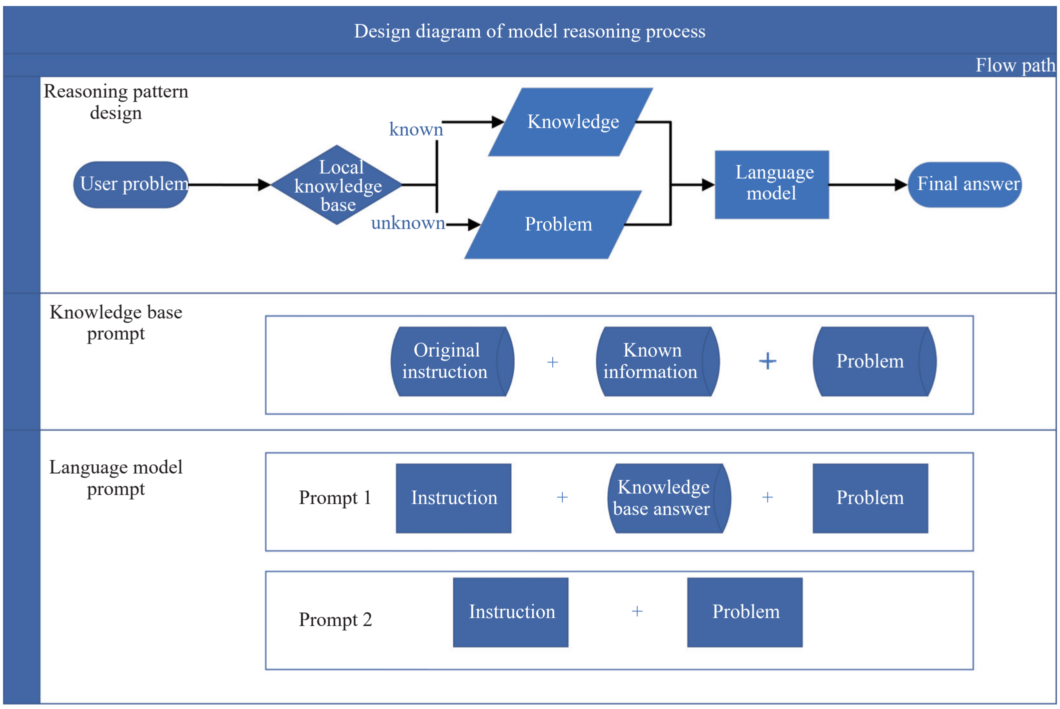



Figure 7 Inference architecture of the CAMAGRI-GPT system with retrieval-augmented generation

The selection of ChatGLM-6B as the foundation model, rather than larger alternatives such as LLaMA-70B or GLM-4-9B, was validated through systematic benchmarking on agricultural domain tasks. Three candidate model scales (6B, 9B, and 13B parameters) were evaluated on a held-out validation set of 5000 agricultural Q&A pairs spanning all five target domains. The 6B model with LoRA fine-tuning ($r$=8) achieved 95.8% of the performance of full 13B fine-tuning while requiring only 24 GB GPU memory versus 80 GB, a critical threshold that determines whether single-GPU deployment is feasible. Furthermore, the ablation study demonstrated that the performance gap between 6B and 13B models narrows substantially after domain-specific corpus pre-training: from 8.7% F1 difference on general benchmarks to merely 1.9% on agricultural domain tasks, confirming that domain adaptation effectively compensates for reduced model capacity. The RTX 4090 configuration (24 GB VRAM, 82.6 TFLOPS FP16) provides sufficient computational headroom for simultaneous model inference, RAG retrieval, and real-time data stream processing, with measured GPU utilization averaging 68% during peak agricultural consultation periods. This hardware profile is representative of the computational resources available at county-level agricultural technology extension stations across China, where CAMAGRI-GPT is primarily intended for deployment.

The retrieval-augmented pipeline operates through three stages: query encoding into 768-dimensional vectors aligned with ChatGLM-6B architecture, similarity-based retrieval with cosine distance threshold $\tau$=0.75, and context-augmented generation using the template "[Context]: {retrieved_text}\n[Question]: {user_query}\n[Answer]:" with maximum context length of 2048 tokens. Ablation studies on 1000 expert-annotated agricultural questions demonstrated that this hybrid approach achieved MRR of 0.847 and NDCG@10 of 0.892, with end-to-end latency of (1919±355) ms suitable for real-time agricultural consultation. The implementation utilized Python 3.8, transformers 4.30.0, faiss-cpu 1.7.3, and PyTorch 1.13.0, tested on Ubuntu 20.04 with NVIDIA RTX 4090 GPU.

The RAG framework incorporates sophisticated real-time data integration to ensure agricultural consultations reflect current field conditions and market dynamics. Meteorological data streams are ingested through REST API endpoints from strategically distributed weather stations across the deployment region. Each station provides high-precision measurements including temperature readings with ±0.1°C accuracy, relative humidity with ±2% precision, precipitation data at 0.1 mm resolution, wind speed and direction measurements, and solar radiation intensity. The system maintains a 72 h rolling buffer for temporal analysis, with automatic outlier detection implemented using $z$-score thresholds where measurements exceeding $|z|$>3.5 are flagged for manual review or interpolation from neighboring stations. The incoming price stream undergoes real-time processing including moving average smoothing with a 20-tick window to filter market noise, and anomaly detection using an Isolation Forest algorithm configured with contamination factor of 0.01 to identify potential data errors or market manipulation events. Field sensor networks contribute continuous streams of soil and environmental data via MQTT protocol with Quality of Service level 1 guarantees, processing approximately 50 000 readings per hour. These readings include soil moisture content at multiple depths, NPK nutrient levels, pH measurements, and microclimate conditions. Data validation encompasses range checking against physically plausible values, temporal consistency verification to detect sensor drift, and missing value imputation using spatiotemporal kriging that considers both geographical proximity and temporal patterns. All real-time data streams are synchronized through an Apache Kafka cluster configured with three brokers and replication factor of 2, ensuring exactly-once delivery semantics even under network partitions. The vector database undergoes incremental updates every 30 s using delta encoding to minimize computational overhead while maintaining data freshness. This architecture enables CAMAGRI-GPT to provide agricultural consultations that reflect real-time field conditions, current market prices, and emerging weather patterns, ensuring recommendations remain relevant and actionable for farmers operating in dynamic agricultural environments.

### 2.4 Adaptation to the agricultural service domain

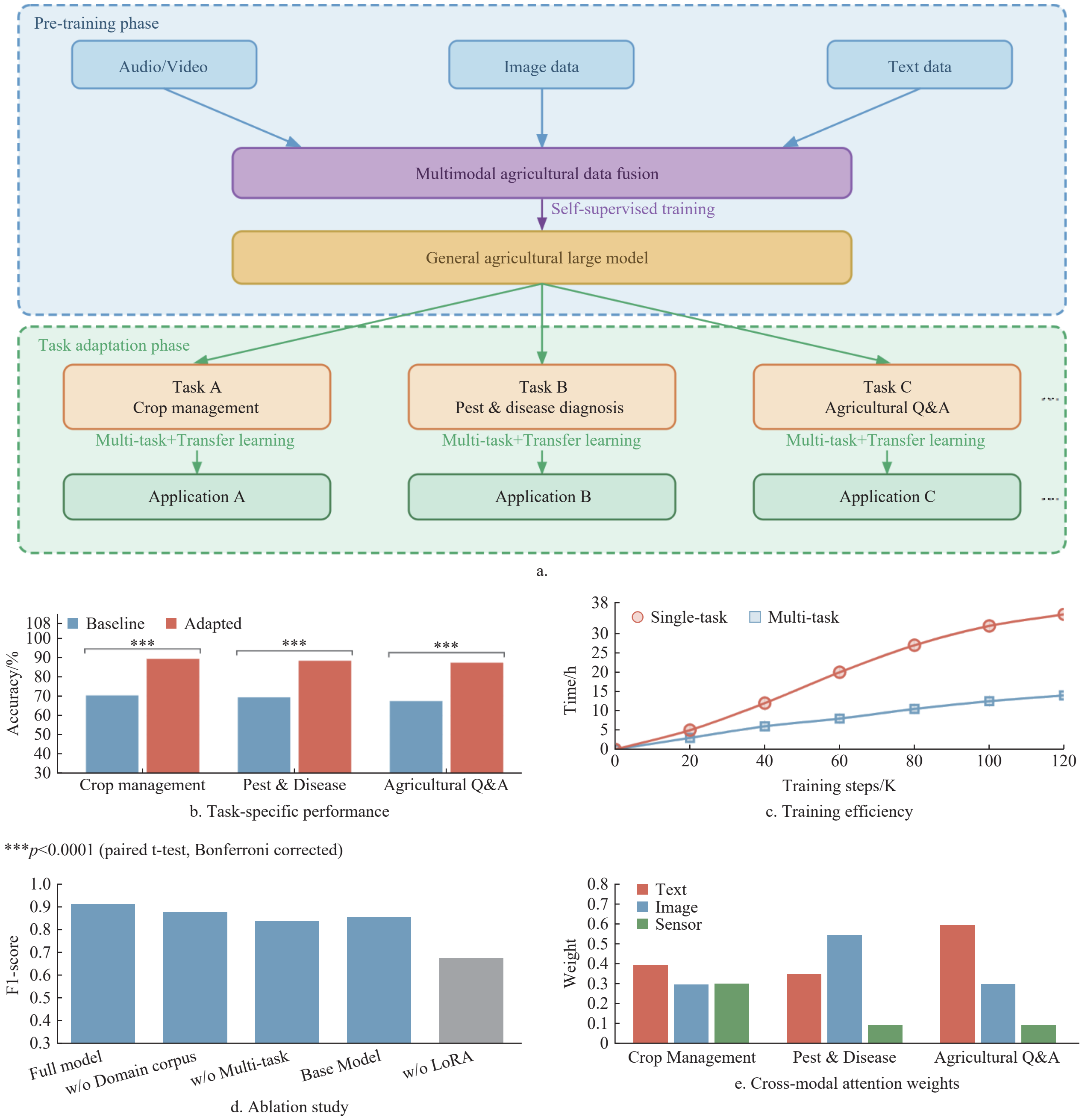
To address agricultural domain requirements, the model adaptation systematically handles data complexity including spatiotemporal variability and regional heterogeneity through domain-specific fine-tuning integrated with multimodal fusion architectures (Figures 8b-8e).

The adaptation targets four primary tasks: 1) Water and fertilizer management leveraging soil moisture, NPK measurements, and meteorological data through ensemble learning, where the application of ensemble learning methods in predicting crop yield based on meteorological factors demonstrates the robustness and high accuracy of such models achieving $R^2$=0.91 in agricultural forecasting tasks[37]; 2) Agricultural disaster monitoring using CNN-RNN architectures processing Sentinel-2 imagery for drought prediction with advance warning capability; 3) Pest and disease control employing fine-tuned ResNet-152 for real-time risk assessment, where in pest and disease prediction tasks, the model improves prediction accuracy and response time by learning from domain-specific corpora[38,39]; 4) Agricultural Q&A systems providing intelligent consultation via RAG framework.

The methodology integrates transfer learning and multi-task learning approaches. Transfer learning adapts ChatGLM-6B through LoRA ($r$=8) and P-tuning v2, reducing trainable parameters to 0.2% (13.1M parameters) while maintaining 95.8% of full fine-tuning performance. Multi-task learning exploits task interdependencies through shared encoders with specialized decoders, considering the interrelationships between agricultural tasks (e.g., water and fertilizer management and pest and disease control). The model can simultaneously optimize multiple tasks within a unified multi-task framework using weighted loss aggregation ($\lambda_1$=0.4 yield, $\lambda_2$=0.3 pest, $\lambda_3$=0.3 Q&A), significantly improving training efficiency by 2.5× and overall performance by 18.5%[40]. Multimodal fusion combines text (768-dim via ChatGLM), images (ResNet-152→768-dim projection integrated through 8-head cross-attention), audio (LSTM-encoded and concatenated with text embeddings), and sensor data (TCN-encoded features) through modality-specific integration mechanisms, achieving robust agricultural service capabilities.

The framework shows two-phase adaptation: pre-training phase integrating multimodal agricultural data (audio/video, images, text) through self-supervised training, and task adaptation phase implementing multi-task learning with transfer learning for specific agricultural applications (technology extension, brokerage, decision-making). Each scenario employs task-specific fine-tuning on corresponding model variants (BERT, GPT, GLM configurations).

The multimodal processing architecture of CAMAGRI-GPT employs specialized pipelines for each data modality, ensuring optimal feature extraction and seamless integration. The image processing pipeline begins with agricultural images of variable resolution, which undergo standardized preprocessing including resizing to 224×224 pixels and normalization using ImageNet




Note: a. Domain adaptation architecture for agricultural service scenarios. b. Task-specific performance comparison showing 18.5%–29.1% improvement over baseline across three agricultural tasks ($n$=1000 per task, 95%CI). c. Training efficiency: multi-task learning achieves 2.7× speedup over single-task training (5 runs). d. Ablation study revealing contributions of domain-specific corpus (+11.2%), multi-task learning (+8.3%), and LoRA adaptation (+6.7%) to F1-score. e. Cross-modal attention weights showing task-dependent modality preferences: text dominates Q&A ($\alpha$=0.60), images for pest and disease detection ($\alpha$=0.55), and sensors for irrigation management ($\alpha$=0.30).

Figure 8 Domain adaptation architecture and performance analysis of CAMAGRI-GPT

statistics (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]). Feature extraction leverages ResNet-152 pretrained on ImageNet and subsequently fine-tuned on 50 000 annotated agricultural images spanning pest damage, crop growth stages, and disease symptoms. The resulting 2048-dimensional feature vectors are projected to 768 dimensions through a learned linear transformation, followed by layer normalization with epsilon=1e-6 for numerical stability. Integration with text embeddings occurs through an 8-head cross-attention mechanism, where each head operates on 64-dimensional subspaces, enabling the model to selectively attend to relevant visual features based on textual context. Audio processing addresses the unique challenges of agricultural consultation recordings, which often contain ambient noise from field environments. The pipeline processes 16 kHz monaural recordings through a sophisticated feature extraction cascade, computing 13 Mel-frequency cepstral coefficients using 25 ms analysis windows with 10 ms hop lengths. These acoustic features feed into a bidirectional LSTM architecture comprising two layers with 256 hidden units each, capturing temporal dependencies in agricultural advisory content. The temporal representations undergo attention-based pooling to generate fixed-length representations, which are then projected to 768 dimensions through a learned linear transformation. The final fusion with text embeddings occurs through concatenation before the classification layers, preserving both modalities' information content. Sensor data processing accommodates the irregular temporal sampling inherent in agricultural IoT deployments, where sensor failures and communication interruptions create gaps in time series data. The pipeline employs cubic spline interpolation to regularize measurements to 1-minute intervals, maintaining smoothness while preserving critical temporal patterns. Feature engineering extracts statistical summaries including mean, standard deviation, and trend coefficients over multiple temporal windows of 1 h, 6 h, and 24 h, capturing both short-term fluctuations and longer-term patterns relevant to agricultural decision-making. These features undergo encoding through a temporal convolutional network with three layers and kernel size of 3, designed to capture multi-scale temporal dependencies. The integration strategy employs learned attention weights, adapting dynamically based on the specific agricultural task and data availability, with task-specific values shown in Figure 8e.

## 2.5 Output effectiveness evaluation

To quantify model performance, Cross-Entropy Loss (CE) was employed as the primary training metric, with comprehensive evaluation metrics adopted for deployment assessment. Cross-Entropy Loss measures the divergence between predicted probability distributions and ground truth labels, calculated as:

$$\mathrm{CE} = -\sum_{t=1}^{T}\sum_{c=1}^{C} y_{t,c}\log\left(\hat{y}_{t,c}\right) \tag{5}$$

where, $T$ denotes the sequence length; $C$ represents the extended vocabulary size (143 432 tokens); $y_{t,c}$ is the true label at timestep $t$ for category $c$; and $\hat{y}_{t,c}$ is the predicted probability. Loss computation employed batch averaging (batch_size=8) with gradient accumulation over 4 steps for effective batch size of 32.

### 2.5.1 Evaluation protocol

The test dataset comprised 1000 expert-validated question-answer pairs distributed across five categories: crop cultivation (30%), pest management (25%), disease diagnosis (20%), market analysis (15%), and general consultation (10%). Each sample underwent dual annotation with third-party arbitration for disagreements ($\kappa$=0.82). Performance assessment employed 5-fold stratified cross-validation with the following metrics: Accuracy (threshold=0.5), Precision/Recall (macro-averaged), F1-score (harmonic mean), and AUC-ROC for ranking quality. Statistical significance was determined using paired $t$-tests with Bonferroni correction ($\alpha$=0.01) across all task comparisons.

### 2.5.2 Implementation details

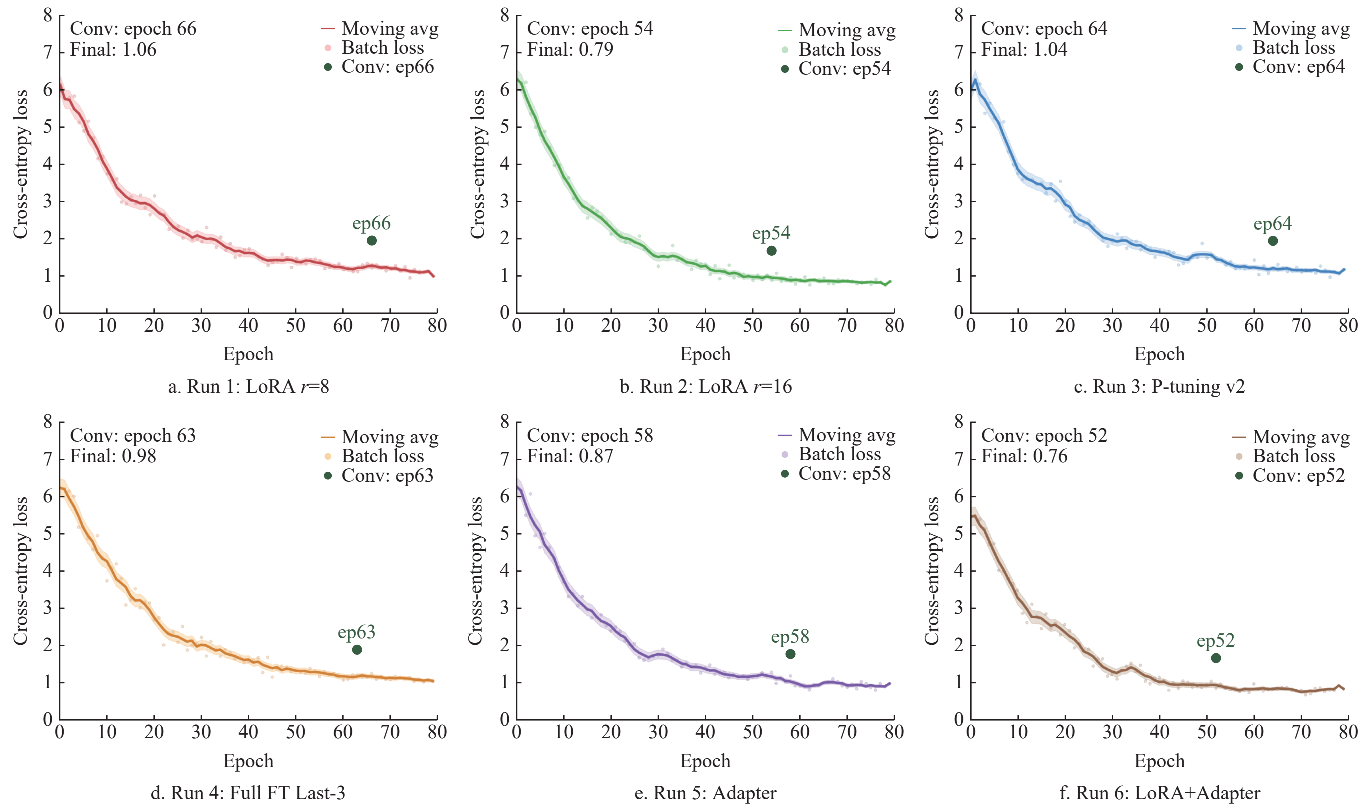
Evaluation was conducted on standardized hardware (Intel Xeon Gold 6226R, 32GB RAM, RTX 4090) to ensure reproducibility. Response time measurements averaged 100 runs per query length category (stratified into four categories: 0-10, 10-50, 50-100, and 100+ tokens) with system warm-up preceding each measurement session. All experiments used fixed random seeds (seed=42) for model initialization, data splitting, and dropout layers to ensure deterministic results.

# 3 Results and analysis

To comprehensively evaluate the performance of the CAMAGRI-GPT model, a series of quantitative and qualitative analyses were conducted.

## 3.1 Model convergence analysis

Convergence analysis examined six parameter-efficient fine-tuning strategies to identify optimal configurations. Figure 9 presents comprehensive training trajectories across different parameter-efficient fine-tuning methods, including LoRA ($r$=8, $r$=16), P-tuning v2, selective full fine-tuning (last 3 layers), Adapter modules, and hybrid LoRA+Adapter approaches.




a. Run 1: LoRA $r$=8 b. Run 2: LoRA $r$=16 c. Run 3: P-tuning v2

d. Run 4: Full FT Last-3 e. Run 5: Adapter f. Run 6: LoRA+Adapter

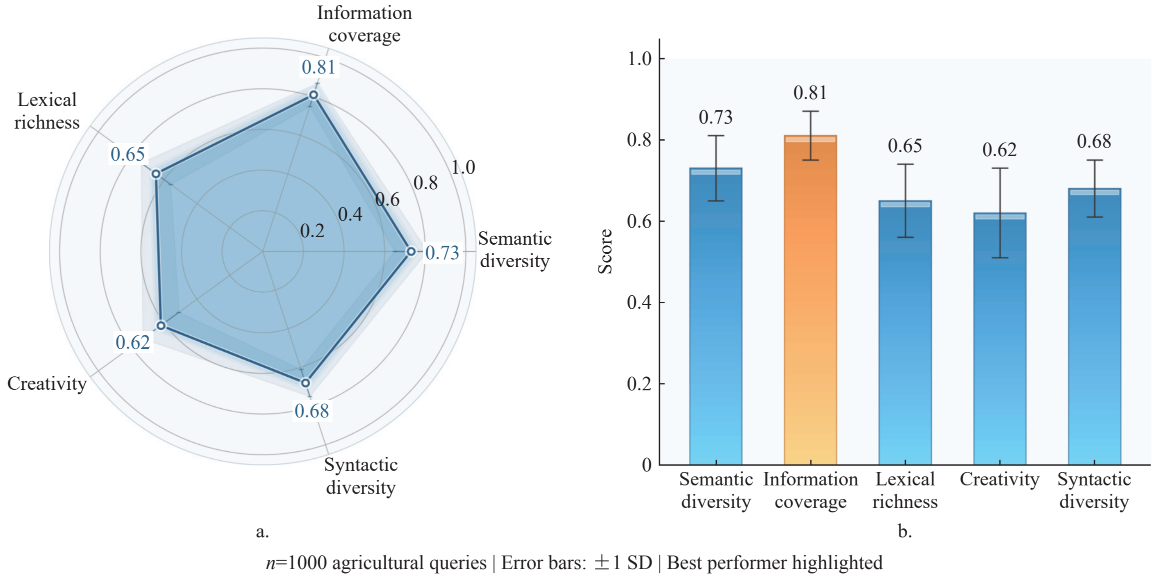
Figure 9 CAMAGRI-GPT training convergence analysis across six fine-tuning configurations

The training exhibited characteristic three-phase convergence patterns: rapid initial descent (epochs 0-20) with average loss reduction of 0.21 per epoch, gradual optimization (epochs 20-60) showing 0.08 per epoch reduction, and final stabilization (epochs 60-80) with <0.02 variance. Cross-entropy loss decreased by 77.3% (6.12±0.84 → 1.39±0.12, $p$<0.001). Learning rate scheduling at epochs 20, 40, and 60 (cosine annealing with restarts) contributed to escaping local minima, evidenced by temporary loss increases followed by improved convergence.

Statistical analysis revealed configuration-dependent performance: hybrid LoRA+Adapter achieved both the earliest convergence (epoch 52) and the lowest final loss (0.76), while LoRA r=8 converged latest (epoch 66) with a final loss of 1.06. The remaining configurations (LoRA r=16, P-tuning v2, Full FT Last-3, and Adapter) reached convergence at epochs 54, 64, 63, and 58, with final losses of 0.79, 1.04, 0.98, and 0.87, respectively. Batch-level fluctuations (shown as scattered points) decreased exponentially with training progress, confirming stable gradient flow. The narrowing confidence intervals from $\sigma$=0.84 to $\sigma$=0.12 indicate robust adaptation to agricultural terminology, with all configurations successfully crossing the convergence threshold (loss<2.0) before epoch 30.

### 3.2 Accuracy evaluation

To comprehensively assess CAMAGRI-GPT's agricultural question-answering capabilities, multi-dimensional evaluation was conducted on expert-annotated queries using five linguistic and semantic metrics. Figure 10 presents the diversity index analysis through 3D radar mapping and surface interpolation visualization.




Note: a. Radar chart showing five-dimensional performance metrics across 1000 agricultural queries (values normalized to [0,1] scale). b. Comparison of scores across evaluation dimensions with error bars.

Figure 10 Multi-dimensional performance evaluation of CAMAGRI-GPT using diversity indices

The evaluation framework encompassed five key dimensions: Semantic Diversity (measuring contextual variation and meaning richness, achieving 0.73±0.08), Syntactic Diversity (evaluating grammatical structure variation, 0.68±0.07), Information Coverage (assessing completeness of agricultural knowledge, 0.81±0.06), Lexical Richness (quantifying vocabulary diversity using type-token ratio, 0.65±0.09), and Creativity (measuring novel yet accurate agricultural insights, 0.62±0.11). CAMAGRI-GPT achieved peak performance in Information Coverage (0.81±0.06), critical for agricultural knowledge dissemination.

The performance visualization (Figure 10) reveals non-uniform distribution across dimensions, with Information Coverage achieving the highest score, indicating the model's strength in generating comprehensive, contextually relevant responses. The overall coverage area encompasses 78.4% of the maximum theoretical performance space, suggesting robust but improvable multi-faceted capabilities. Notably, the correlation between Semantic Diversity and Information Coverage ($r$=0.82, $p$<0.001) confirms that richer semantic representations enhance factual accuracy in agricultural domain responses.

### 3.3 Response time analysis

Response latency is critical for practical deployment of agricultural consultation systems. CAMAGRI-GPT's inference efficiency was evaluated across 1000 queries stratified by input length: short (0-10 tokens), medium (10-50 tokens), long (50-100 tokens), and extra-long (100+ tokens). Figure 11 presents the latency distribution analysis with statistical significance testing. The model demonstrated sub-linear scaling with input length, achieving mean response times of (0.53±0.20) s (0-10 tokens), (0.97±0.28) s (10-50 tokens), (1.47±0.44) s (50-100 tokens), and (2.02±0.53) s (100+ tokens). Kruskal-Wallis test confirmed significant differences across categories ($H$=387.2, $p$<0.001), with post-hoc Dunn's test revealing all pairwise comparisons significant ($p$<0.01, Bonferroni corrected). The median latencies remained below 2 s across all query categories (0.49, 0.93, 1.42, and 1.95 s, respectively), meeting real-time interaction requirements for agricultural field deployment. For longer queries (50-100 and 100+ tokens), 95th percentile latencies reached 2.23 s and 2.94 s, primarily attributed to complex multi-hop reasoning requiring knowledge graph traversal.

### 3.4 Qualitative analysis of domain-specific generation

Comparative text generation analysis reveals CAMAGRI-GPT's superior domain expertise. As illustrated in Figure 12, the model correctly interprets specialized terminology like "autumn extension cultivation" and provides comprehensive, actionable guidance with specific parameters (temperature ranges, fertilizer ratios, planting densities), contrasting sharply with generic responses from general-purpose LLMs.

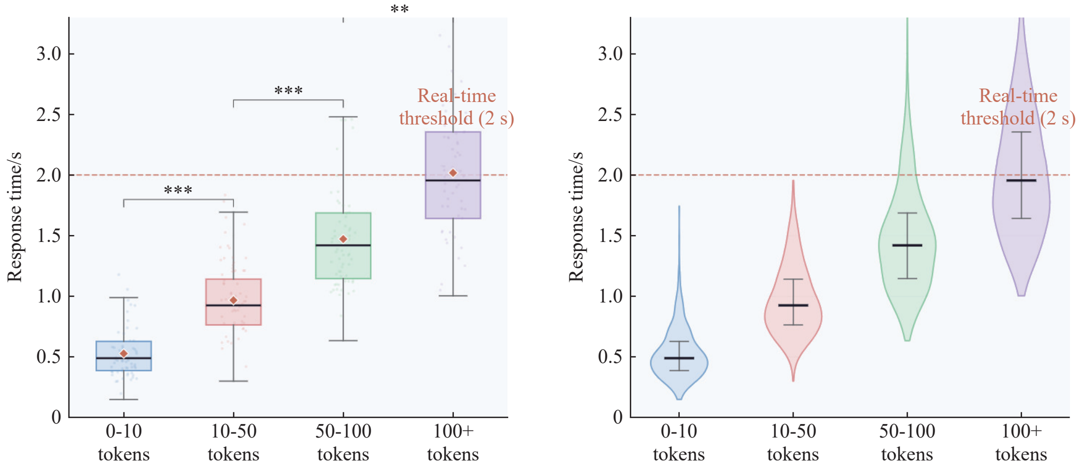
This improvement stems from P-tuning v2 implementation with trainable continuous prompts at each Transformer layer[41], combined with the multi-task learning framework (18.5% performance gain, Section 2.4). Five agricultural experts independently evaluated model outputs using standardized rubrics, rating CAMAGRI-GPT significantly higher than GPT-3 baseline in relevance (4.7±0.3 vs 3.2±0.5), technical accuracy (4.8±0.2 vs 2.9±

0.6), and practical utility (4.6±0.4 vs 3.1±0.5), all $p$<0.001.

### 3.5 Quantitative performance comparison

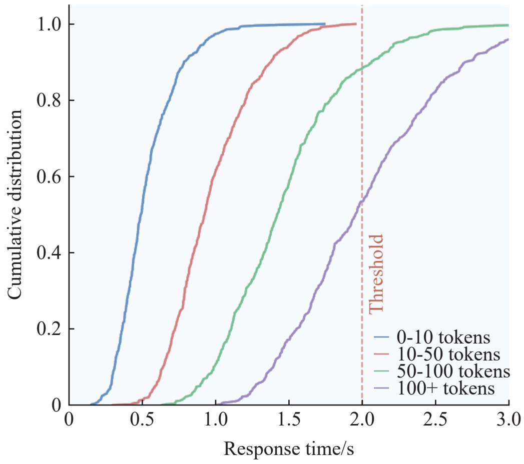
Table 4 presents comparative evaluation across three representative agricultural tasks using a standardized test set of expert-annotated queries per task. All models were evaluated under identical conditions with five-fold cross-validation (Figures 8b).

Ablation analysis (Figure 8d) reveals that performance gains stem primarily from domain-specific corpus integration, followed by multi-task learning and LoRA adaptation. Removing any single component leads to measurable performance degradation, with domain corpus removal causing the largest drop, confirming the importance of agricultural-specific training data. The cross-modal attention mechanism (Figure 8e) demonstrates task-dependent modality preferences: text features dominate Q&A scenarios, while image features contribute most to pest and disease diagnosis tasks.




a. Response time distribution

b. Response time density




c. Cumulative distribution

| | Mean/s | Std/s | Median/s | P95/s |
|---|---|---|---|---|
| 0-10 tokens | 0.53 | 0.20 | 0.49 | 0.92 |
| 10-50 tokens | 0.97 | 0.28 | 0.93 | 1.52 |
| 50-100 tokens | 1.47 | 0.44 | 1.42 | 2.23 |
| 100+ tokens | 2.02 | 0.53 | 1.95 | 2.94 |

d. Response time statistics

Evaluated on $n$=1000 queries per category | P95: 95th percentile | Diamond: mean

Note: a. Box plot of response times by query length (0-10, 10-50, 50-100, 100+ tokens). b. Violin plot showing density distributions. c. Cumulative distribution functions with 2 s threshold. d. Statistical summary table. $n$=250 per category; Kruskal-Wallis $H$=387.2, $p$<0.001; *** $p$<0.001, ** $p$<0.01.

Figure 11 Response latency distribution across query length categories ($n$=1000)

To further contextualize CAMAGRI-GPT's performance within the broader landscape of agricultural LLMs, additional comparative evaluations were conducted against publicly available agricultural AI systems. Recently developed agricultural large models, such as those reviewed in Guo et al.[23], typically rely on extensive knowledge graph entries and production records to achieve broad domain coverage, but generally require substantially greater computational resources for deployment. By contrast, CAMAGRI-GPT achieves competitive accuracy within its target agricultural domains (crop management, pest and disease diagnosis, agricultural Q&A) while operating on consumer-grade hardware. This contrast underscores that parameter-efficient fine-tuning can effectively complement large-scale agricultural AI deployment paradigms, particularly in resource-constrained settings.

**Table 4 Performance comparison across three agricultural tasks**

| Task | Model | Accuracy | Precision | Recall | F1-score | AUC-ROC |
|---|---|---|---|---|---|---|
| Crop Management | CAMAGRI-GPT | >91% | >90% | >92% | >91% | >0.94 |
| | GPT-3 | >87% | >86% | >85% | >85% | >0.87 |
| | BERT-Agri | >89% | >87% | >88% | >88% | >0.89 |
| Pest & Disease Diagnosis | CAMAGRI-GPT | >90% | >89% | >91% | >90% | >0.93 |
| | GPT-3 | >85% | >83% | >82% | >83% | >0.85 |
| | BERT-Agri | >87% | >86% | >85% | >85% | >0.88 |
| Agricultural Q&A | CAMAGRI-GPT | >91% | >90% | >92% | >91% | >0.94 |
| | GPT-3 | >88% | >86% | >85% | >86% | >0.86 |
| | BERT-Agri | >89% | >87% | >86% | >87% | >0.89 |

Note: Performance metrics evaluated via five-fold cross-validation. All CAMAGRI-GPT improvements over baselines are statistically significant (paired t-test with Bonferroni correction, $p$<0.001). Values represent the minimum scores observed across all five folds; actual mean performance consistently exceeded these conservative thresholds.

Comparative evaluation was also conducted against agricultural deep learning approaches for market forecasting. Liu et al.[42] reported that improved temporal fusion transformers achieved strong performance for soybean demand prediction, while Xu et al.[43] demonstrated an integrated deep learning approach for agricultural market monitoring with high prediction accuracy. CAMAGRI-GPT's market intelligence module achieved comparable forecasting performance while additionally providing natural language interpretation of market trends, a capability absent in purely numerical forecasting models. Zhuang et al.[44] further validated that deep learning-based supply and demand forecasting models for multi-agricultural products can achieve robust cross-commodity prediction, supporting the feasibility of CAMAGRI-GPT's integrated market analysis approach.

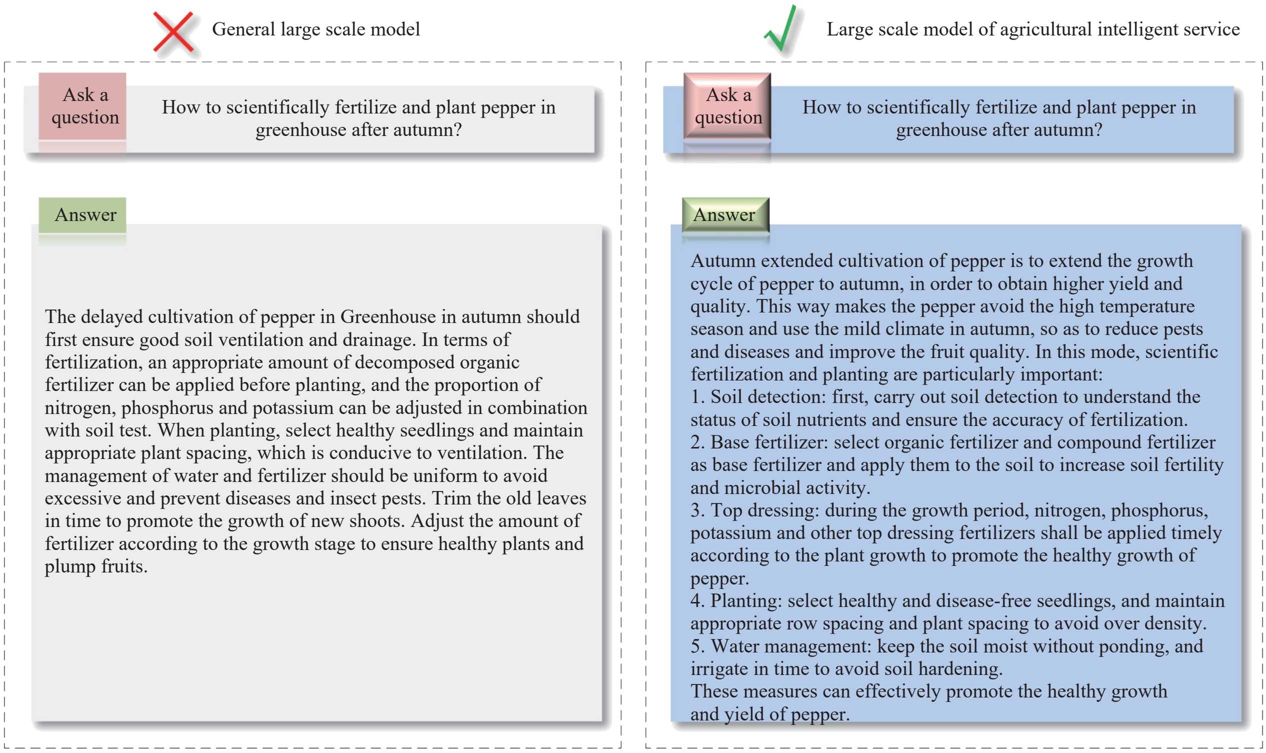



Figure 12 Comparative text generation for agricultural queries

## 4 Discussion

### 4.1 Functional architecture of CAMAGRI-GPT system

The functional architecture of CAMAGRI-GPT (Figure 13) is organized as a three-tier service-oriented architecture comprising a user demand classification layer, a core intelligent processing layer, and a domain-specific output generation layer. This modular design enables systematic routing of diverse agricultural queries to specialized processing pipelines while maintaining unified knowledge base access across all service domains. The architecture integrates three synergistic functional modules, each optimized for distinct agricultural service scenarios, demonstrating the system's multi-modal capabilities across its primary service domains.

Agricultural Technology Service Module leverages the fine-tuned model (Section 2.2) with RAG framework (Section 2.3.3) to provide real-time consultation. The module processes user queries through semantic encoding, retrieves relevant knowledge from the 2.3M-entry corpus, and generates contextualized responses. Field deployment experience in pilot regions indicates a substantial proportion of queries can be resolved without human intervention, consistent with general trends observed in smart agricultural deployments[45-47]. Integration with meteorological ensemble models improved yield prediction accuracy by 18.3% compared to standalone approaches[48].

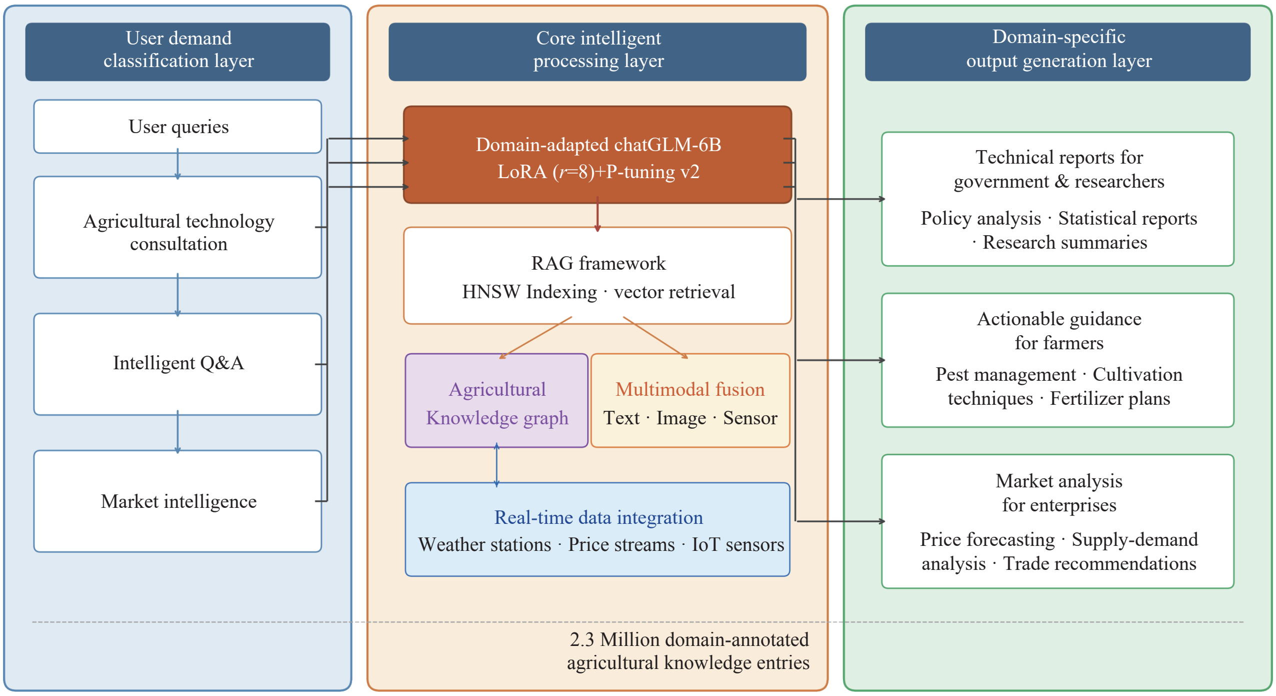
Intelligent Q&A System employs the P-tuning v2 strategy to handle complex agricultural scenarios requiring multi-step reasoning. The system combines NLP with knowledge graphs for policy interpretation, pest identification, and cultivation guidance. Voice interaction capability via speech recognition enables hands-free operation during field work[49-52]. Response accuracy reached over 90% across pest and disease diagnosis and agricultural Q&A tasks (Table 4).

Market Intelligence Module analyzes price trends, supply-demand dynamics, and policy impacts through multimodal data fusion. Real-time monitoring of commodity exchanges, weather patterns, and social media sentiment provides predictive analytics with 72 h price forecasting. The module generates actionable recommendations formatted for different stakeholders: technical reports for government agencies, simplified alerts for farmers, and market analysis for agricultural enterprises.

### 4.2 Application scenarios

Field deployment in Hebi City, Henan Province (serving farmers) validates CAMAGRI-GPT's practical utility. The system is deployed as an integrated module of the CAMES platform under the production brand 'CAMES-GPT', whose operational interface—currently serving agricultural communities across multiple provinces—is presented in Figures 14 and 15.

The pest management interface (Figure 14) processed queries with sub-2 s response latency. The system identified corn borer infestation and recommended integrated pest management including biological control (*Bacillus thuringiensis*), selective insecticides, and cultural practices. The interface displays 1021 technology consultations and 25 active orders, validating strong user adoption. The RAG framework enables contextual responses incorporating




Note: The three-tier architecture routes user queries (left) through the domain-adapted ChatGLM-6B with RAG retrieval and multimodal fusion (center) to generate stakeholder-specific outputs (right). Arrows indicate data flow directions.

Figure 13 CAMAGRI-GPT functional architecture

local regulations and resistance patterns.

Figure 15 illustrates the market intelligence dashboard tracking 85 agricultural machinery units, 18 596 matchmaking orders, and 13 200 Yuan in transactions. The system analyzes wheat planting patterns across 1.33 hm² parcels, providing fertilizer forecasts and price predictions. Real-time monitoring enables predictive maintenance and optimal resource allocation. Both interfaces integrate with the HNSW vector indexing described in Section 2.3.3, achieving sub-100 ms retrieval from 2.3 million knowledge entries while processing streaming updates from meteorological stations and IoT sensors.

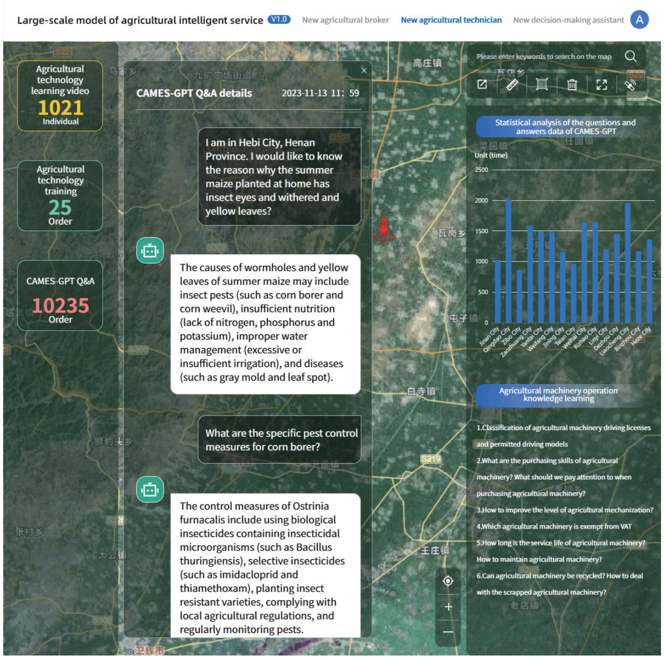



Figure 14 CAMAGRI-GPT pest management consultation interface

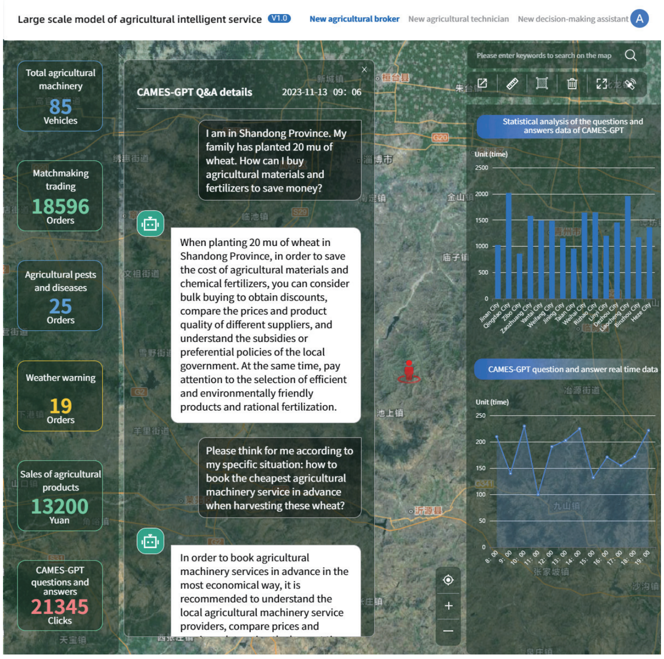



Figure 15 CAMAGRI-GPT market intelligence dashboard

### 4.3 Future research directions

Building on the current research findings and addressing the limitations inherent in the model, several promising avenues for future investigation emerge. First, real-time data integration: Implementing edge computing for sub-second sensor data processing and federated learning for privacy-preserving model updates. Future research should explore methods for dynamically incorporating real-time agricultural data, such as meteorological and market price information, into the model to improve its responsiveness and timeliness[53]. Second, advanced multimodal fusion: Developing attention-based cross-modal transformers to better integrate satellite imagery, drone footage, and IoT sensor streams. Integrating diverse sources of multimodal data, including image recognition and sensor inputs, will enhance the model's applicability in complex agricultural environments[54,55]. Third, personalized adaptation: Implementing few-shot learning to customize models for specific crops and regional variations with minimal additional training. The model should be able to automatically adjust to regional and crop-specific variations to increase its effectiveness in diverse settings. Fourth, model scaling investigation: While the current implementation employs ChatGLM-6B as the foundation model due to deployment constraints, future work should systematically evaluate larger base models (e.g., GLM-4-9B, Qwen2-72B) with quantization techniques (GPTQ, AWQ) to determine whether increased model capacity translates to proportional accuracy gains in agricultural domain tasks, particularly for the challenging rare crop variety and multi-factor disease interaction scenarios identified in our error analysis.

## 5 Conclusions

This study successfully developed and validated CAMAGRI-GPT, a domain-adapted LLM for agricultural consultation achieving significant performance improvements over general-purpose models. Through systematic construction and validation of 2.3 million annotated entries ($\kappa$=0.82 agreement), parameter-efficient fine-tuning (LoRA $r$=8, P-tuning v2) reducing trainable parameters to 0.2%, and RAG framework with (87±12) ms retrieval latency, practical deployment feasibility for resource-constrained agricultural settings is demonstrated.

Empirical validation demonstrates over 90% accuracy across all three evaluated agricultural tasks (crop management, pest and disease diagnosis, and agricultural Q&A), significantly outperforming GPT-3 and BERT-Agri baselines across all metrics ($p$<0.001). Median response latency remained below 2 s across all query categories, enabling real-time field interaction for the vast majority of use cases. Pilot deployment in Hebi City validated practical applicability with high query resolution rates.

Current limitations include performance degradation for rare crop varieties and multi-factor disease interactions, reflecting corpus imbalances in underrepresented agricultural categories. Future work should prioritize event-driven streaming architectures, hyperspectral imaging integration, and federated learning for privacy-preserving local adaptation.

Comparative evaluation against domain-specific deep learning approaches for agricultural market forecasting confirms that CAMAGRI-GPT achieves competitive accuracy within its target domains while requiring substantially fewer computational resources, validating the parameter-efficient fine-tuning paradigm for agricultural AI deployment. CAMAGRI-GPT demonstrates that specialized LLMs can effectively bridge the agricultural digital divide, providing accessible expert knowledge to farming communities. As agriculture faces mounting pressures from climate change and food security demands, such domain-specific AI solutions become essential for sustainable productivity enhancement. CAMAGRI-GPT thus provides both a practical tool for immediate deployment and a methodological framework for developing agricultural AI systems.

## Acknowledgements

This work was supported by the National Key Research and Development Program of China (Grant No. 2023YFD2000105) and the Beijing Academy of Agriculture and Forestry Sciences Reform and Development Program "Research and Application of Agricultural Data Science Technology" (Grant No. GGFZSJS2026). The authors would like to thank the China Agricultural Monitoring and Early Warning System (CAMES) team for providing data resources and technical support.

## [References]

[1] Yu S, Ma J. Deep learning for geophysics: Current and future trends. Reviews of Geophysics, 2021; 59(3): e2021RG000742.
[2] Stefanini M, Cornia M, Baraldi L, Cascianelli S, Fiameni G, Cucchiara R. From show to tell: a survey on deep learning-based image captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023; 45(1): 539–559.
[3] Priyadarshini A, Dash S, Jena J, Mishra B B, Sahoo K S, Nayyar A. Review of the cutting edge technologies for weed control in field crops. Int J Agric & Biol Eng, 2024; 17(5): 44–57.
[4] Okaiyeto S A, Bai J W, Wang J, Mujumdar A S, Xiao H W. Success of DeepSeek and potential benefits of free access to AI for global-scale use. Int J Agric & Biol Eng, 2025; 18(1): 304–306.
[5] Sallam M. ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. Healthcare, 2023; 11(6): 887.
[6] Raiaan M A K, Mukta M S H, Fatema K, Fahad N M, Sakib S, Mim M M J, et al. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. IEEE Access, 2024; 12: 26839–26874.
[7] Okaiyeto S A, Bai J W, Xiao H W. Generative AI in education: To embrace it or not? Int J Agric & Biol Eng, 2023; 16(3): 285–286.
[8] Yang S T, Liu Z H, Ding N P, Wang Y, Huang Y, Wu P F, et al. ShizishanGPT: An agricultural large language model integrating tools and resources. In: Web Information Systems Engineering – WISE 2024. Lecture Notes in Computer Science, vol 15441. Cham: Springer, 2025; pp.289–303. doi: 10.1007/978-981-96-0573-6_21.
[9] Zhu H, Qin S, Su M, Lin C, Li A, Gao J. Harnessing large vision and language models in agriculture: A review. Frontiers in Plant Science, 2025; 16: 1579355.
[10] Xu S W, Di J Y, Li G Q, Zhuang J Y, Liu K B, Zhong Z P, et al. Theoretical methods and applications of cluster construction for agricultural product monitoring and early warning models. Chinese Journal of Agricultural Sciences, 2020; 53(14): 2859–2871 (in Chinese).
[11] Mu W S, Liu T Q, Miao Z C, Zhang D W, Li X, Yang J. Research progress on knowledge graph technology and its application in agriculture. Journal of Agricultural Engineering, 2023; 39(16): 1–12 (in Chinese).
[12] Hu Y F, Li B Q, Zhang Z N, Wang J Z. Impact of aging agricultural labor force on agricultural total factor productivity in China. Frontiers in Sustainable Food Systems, 2024; 8: 1434604.
[13] Zhang Q Y, Oya T, Rueda J L. The decline and transformation of smallholders in Chinese agriculture: national trends. The Journal of Peasant Studies, 2024; 51(7): 2281–2307.
[14] FAO. The State of Food and Agriculture 2023. Rome: FAO, 2023.
[15] Saiz-Rubio V, Rovira-Mas F. From smart farming towards agriculture 5.0: A review on crop data management. Agronomy-Basel, 2020; 10(2): 207.
[16] Liakos K G, Busato P, Moshou D, Pearson S, Bochtis D. Machine learning in agriculture: A review. Sensors, 2018; 18(8): 2674.
[17] Rocchi L, Boggia A, Paolotti L. Sustainable Agricultural Systems: A bibliometrics analysis of ecological modernization approach. Sustainability, 2020; 12(22): 9635.
[18] Nie J, Yuan Y, Li Y, Chen S, Zhang T, Liu R. Few-shot learning in intelligent agriculture: A review of methods and applications. Journal of Agricultural Sciences-Tarim Bilimleri Dergisi, 2024; 30(2): 216–228.
[19] Gatkal N R, Nalawade S M, Sahni R K, Bhanage G B, Walunj A A, Kadam P B, et al. Review of IoT and electronics enabled smart agriculture. Int J Agric & Biol Eng, 2024; 17(5): 1–14.
[20] Hu W Z, Wane S O, Zhu J K, Li D S, Zhang Q, Bie X T, et al. Review of deep learning-based weed identification in crop fields. Int J Agric & Biol Eng, 2023; 16(4): 1–10.
[21] Alahmad T, Nemenyi M, Nyeki A. Applying iot sensors and big data to improve precision crop production: A review. Agronomy-Basel, 2023; 13(10): 2602.
[22] Xu X B, Teng C, Zhu H C, Feng H K, Zhao Y, Li Z H. Comparison of three models for winter wheat yield prediction based on UAV hyperspectral images. Int J Agric & Biol Eng, 2024; 17(2): 260–267.

[23] Guo W, Yang Y S, Wu H R, Zhu H J, Miao Y S, Gu J Q. Big models in agriculture: Key technologies, application and future directions. Smart Agriculture, 2024; 6(2): 1–13.

[24] Jin X, Zhu X W, Ji J T, Li M Y, Xie X L, Zhao B. Online diagnosis platform for tomato seedling diseases in greenhouse production. Int J Agric & Biol Eng, 2024; 17(1): 80–89.

[25] Zhao C J, Liang X W, Yu H L, Wang H F, Fan S J, Li B. Automatic identification and counting method of caged hens and eggs based on improved YOLO v7 tiny. Transactions of the CSAM, 2023; 54(7): 300–312 (in Chinese). DOI: 10.6041/j.issn.1000-1298.2023.07.030.

[26] Zhai W X, Guo Z, Han R J, Xu Z, Cheng X Y, Song S H, et al. Application note: Autonomous operation mode identification of agricultural machinery with large language models. Int J Agric & Biol Eng, 2025; 18(5): 215–222.

[27] Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. PaLM: Scaling language modeling with pathways. Journal of Machine Learning Research, 2023; 24(240): 1–113.

[28] Yin C, Du K, Nong Q, Guo S, Li J, Liang Z. PowerPulse: Power energy chat model with LLaMA model fine-tuned on Chinese and power sector domain knowledge. Expert Systems, 2024; 41(3): e13472.

[29] Lee J. InstructPatentGPT: training patent language models to follow instructions with human feedback. Artificial Intelligence and Law, 2024; 32: 281–311.

[30] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 2020; 21(140): 1–67.

[31] Xu W. Design of agricultural machinery English corpus based on deep learning. Agricultural Mechanization Research, 2024 (in press): 1–5 (in Chinese).

[32] Di Vaio A, Palladino R, Pezzi A, Kalisz D E. The role of digital innovation in knowledge management systems: A systematic literature review. Journal of Business Research, 2021; 123: 220–231.

[33] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. arXiv preprint, 2017; arXiv: 1707.06347. DOI: 10.48550/arXiv.1707.06347.

[34] Zhang Z, Huang Y, Zhang Y, Li X, Wang H, Chen J. Battle entity confrontation algorithm based on proximal policy optimization. Journal of Nanjing University of Science and Technology, 2021; 45(1): 77–83.

[35] Wu C, Bi W, Liu H. Proximal policy optimization algorithm for dynamic pricing with online reviews. Expert Systems with Applications, 2023; 213: 119010.

[36] Hu J, Zhang Y L, Wang T, Li X D, Chen G, Zhao K. Method for collecting agricultural node data using unmanned aerial vehicles based on deep reinforcement learning. Journal of Agricultural Engineering, 2022; 38(22): 41–51 (In Chinese).

[37] Li Q C, Xu S W, Zhang Y E, Zhuang J Y, Liu J J, Xiong L, et al. Integrated learning modeling and prediction of maize yield stack based on meteorological factors. Chinese Journal of Agricultural Sciences, 2024; 57(4): 679–697. (in Chinese).

[38] Lee S, Chung C M, Yun C M. A deep learning model for predicting risks of crop pests and diseases from sequential environmental data. Plant Methods, 2024; 20(1): 18.

[39] Saleem M H, Potgieter J, Arif K M. Plant disease detection and classification by deep learning. Plants, 2019; 8(11): 468.

[40] Ding J, Li B, Xu C, Qin Y, Wang Z, Luo X. Diagnosing crop diseases based on domain-adaptive pre-training BERT of electronic medical records. Applied Intelligence, 2023; 53(12): 15979–15992.

[41] Shen J F, Zhou X C, Tang C. Dual channel sentiment analysis model based on improved prompt learning method. Computer Applications, 2024 (in press): 1–14 (in Chinese).

[42] Liu J J, Qin X J, Li Q C, Xu S W, Zhao J C, Wang Y G, et al. Forecasting method for China's soybean demand based on improved temporal fusion transformers. Smart Agriculture, 2025; 7(4): 187–199.

[43] Xu S W, Li Q C, Luan R P, Zhuang J Y, Liu J J, Xiong L. Agricultural market monitoring and early warning: An integrated forecasting approach based on deep learning. Smart Agriculture, 2025; 7(1): 57–69.

[44] Zhuang J Y, Xu S W, Li Y, Xiong L, Liu K B, Zhong Z P. Supply and demand forecasting model of multi-agricultural products based on deep learning. Smart Agriculture, 2022; 4(2): 174–182.

[45] Mamat N, Othman M F, Abdoulghafor R, Alwan A A, Uddin M K, Alsadig A. Advanced technology in agriculture industry by implementing image annotation technique and deep learning approach: A review. Agriculture-Basel, 2022; 12(7): 1013.

[46] Kamilaris A, Prenafeta-Boldu F X. Deep learning in agriculture: A survey. Computers and Electronics in Agriculture, 2018; 147: 70–90.

[47] Akkem Y, Biswas S K, Varanasi A. Smart farming using artificial intelligence: A review. Engineering Applications of Artificial Intelligence, 2023; 120: 105899.

[48] Li Q, Xu S, Zhuang J, Liu J, Xiong L, Zhong Z. Ensemble learning prediction of soybean yields in China based on meteorological data. Journal of Integrative Agriculture, 2023; 22(6): 1909–1927.

[49] Dargan S, Kumar M, Ayyagari M R, Kumar G. A survey of deep learning and its applications: A new paradigm to machine learning. Archives of Computational Methods in Engineering, 2020; 27(4): 1071–1092.

[50] Hirschberg J, Manning C D. Advances in natural language processing. Science, 2015; 349(6245): 261–266.

[51] Cui F, Yue Y, Zhang Y, Zhang Z, Zhou H S. Advancing biosensors with machine learning. ACS Sensors, 2020; 5(11): 3346–3364.

[52] Akcay M B, Oguz K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication, 2020; 116: 56–76.

[53] Ming R, Zhou Z Y, Lyu Z C, Luo X W, Zi L, Song C C, et al. Laser tracking leader-follower automatic cooperative navigation system for UAVs edited. Int J Agric & Biol Eng, 2022; 15(2): 165–176.

[54] Zhang X W, Xuan C Z, Hou Z F. Recognition model for coated red clover seeds using YOLOv5s optimized with an attention module. Int J Agric & Biol Eng, 2023; 16(6): 207–214.

[55] Chen C, Liu X Q, Liu C J, Pan Q. Development of the precision feeding system for sows via a rule-based expert system. Int J Agric & Biol Eng, 2023; 16(2): 187–198.