

Knowledge assimilation: Implementing knowledge-guided agricultural large language model

Jingchi Jiang ^{a,b}, Lian Yan ^a, Haifeng Liu ^a, Zhenbo Xia ^a, Haotian Wang ^a, * Yang Yang ^{a,*},
Yi Guan ^a

^a Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China

^b National Key Laboratory of Smart Farm Technologies and Systems, Harbin 150001, China

ARTICLE INFO

Dataset link: <https://github.com/YanPioneer/KALLM>

Keywords:

Large language model
Agricultural knowledge
Agricultural decision-making
Knowledge assimilation
Retrieval augmented

ABSTRACT

Although supervised fine-tuning (SFT) and retrieval-augmented generation (RAG) can help large language models (LLMs) incorporate domain knowledge, they have the following limitations: (1) Data scarcity. There is a severe lack of high-quality data and knowledge bases on dialogue in agriculture. (2) Token-level oversight. Current SFT primarily focuses on fitting general tokens, neglecting agricultural-specific tokens. It leads to omissions of critical information in responses. (3) Sentence-level hurdle. Agricultural queries necessitate sentence-level evidence support from domain knowledge bases, which poses a challenge to precision evidence retrievers. This paper introduces a novel Knowledge-guided Agriculture LLM (KALLM) designed to facilitate multi-task decision-making in agricultural settings. We begin by addressing the data quality issue by establishing an annotation standard and constructing a comprehensive dataset consisting of 220,000 Q&A pairs derived from authoritative agricultural documents. At the token level, we propose a knowledge-coordinated SFT approach that enhances the representation of agriculture-specific tokens by amplifying their significance during the decoding process. At the sentence level, we introduce a self-reflective RAG mechanism based on topic matching to improve the accuracy of evidence retrieval. Experimental results compared with seven competitive open-domain LLMs and the current SFT-RAG pipeline show that our KALLM achieves state-of-the-art performance and is significantly superior to existing generation frameworks in terms of response fluency, accuracy, and domain fidelity.

1. Introduction

Agricultural production is typically distributed across diverse regions, with significant variations in climate, crop varieties, and farming practices. The complexities of crop growth management and pest and disease systems make achieving scientific and precise production decisions exceedingly challenging. Moreover, as agricultural practices evolve, knowledge has become a critical factor in enabling modernized, data-driven decisions and precision operations. Farmers' demand for agricultural knowledge has surged, yet their access to such knowledge remains limited by current channels and tools [1–5].

In recent years, large language models (LLMs) have achieved exceptional performance in natural language understanding and generation, even reaching human-like proficiency in practical tasks such as query-and-answer (Q&A) [6,7], summarization [8,9], and solution generation [10]. This presents a novel research paradigm for integrating

artificial intelligence into agricultural scenarios, consolidating knowledge, and providing precision decision-making guidance and Q&A-based knowledge services for farming practices. However, due to the scarcity of agricultural knowledge, general LLMs frequently generate unfounded and non-factual contexts, known as hallucination [11]. For example, issues like crop variety recommendations often demand knowledge of local farming conditions, precise crop types, and current pest prevalence, which exceeds the capabilities of general LLMs and search engines [12,13]. Moreover, general LLMs struggle to deliver quantitative and professional responses. Fig. 1 illustrates the responses to a crop-specific pest and disease query given by LLMs, search engines, agricultural experts, and our proposed model. It is evident that existing methods fall short in precision, domain consistency (i.e., alignment with domain knowledge), and knowledge coverage, failing to guide practical farming activities.

* Corresponding author.

E-mail addresses: jiangjingchi@hit.edu.cn (J. Jiang), 23b903008@stu.hit.edu.cn (L. Yan), 24S003142@stu.hit.edu.cn (H. Liu), 23S103141@stu.hit.edu.cn (Z. Xia), wanght1998@hit.edu.cn (H. Wang), yangyang_hit@cust.edu.cn (Y. Yang), guanyi@hit.edu.cn (Y. Guan).

| Query What are the most common pests and diseases affecting corn, and how can they be prevented and controlled? | |
|---|--|
| GPT-4 | There are several common pests and diseases affecting corn, and here are some examples along with their prevention and control methods: 1. Corn Borer (<i>Ostrinia nubilalis</i>): Destroying the initial stages of the pest, such as egg masses, is the best primary prevention method. Prevention & Control: Employ biological control such as releasing natural enemies of insects and installing insect traps to monitor and collect adult moths. For chemical control, select insecticides that are relatively safe for humans and livestock. For spraying: 2. Southern Corn Rootworm (<i>Pectinotarsus polysrys</i>): Description: A fungal disease characterized by the formation of orange-red spots spores primarily on leaves. Prevention & Control: Promptly apply fungicides containing triazole or thiophanate-methyl to combat the disease. 3. Northern Corn Leaf Blight (<i>Exserohilum turcicum</i>): Description: A fungal disease manifesting as circular or irregular brown lesions on leaves. Prevention & Control: Select resistant varieties; remove infected crop residues; applying fungicides containing triazole or thiophanate-methyl to combat the disease. 4. Gray Leaf Spot (<i>Cercospora zeae-maydis</i>): Description: This disease causes gray or grayish-white spots to appear on leaves. Prevention & Control: Remove crop residues promptly and apply fungicides early in the season to prevent the onset of the disease. |
| CLAUDA | The most prevalent pests and diseases affecting corn, along with their prevention and control methods, are as follows: 1. Corn Borer (<i>Ostrinia nubilalis</i>): Prevention & Control: Timely removal of weeds and crop residues from the field; employing sex pheromones to lure and eliminate adult moths; spraying insecticides such as chlordane-parathion and indoxacarb at appropriate intervals. 2. Northern Corn Leaf Blight (<i>Exserohilum turcicum</i>): Prevention & Control: Selecting disease-resistant varieties; planting at an optimal density to ensure proper ventilation; applying fungicides like propiconazole and azoxystrobin during the initial stages of infection. 3. Corn Rootworm : Prevention & Control: Selecting resistant varieties; removing crop residues; applying fungicides containing triazole or thiophanate-methyl to combat the disease. 4. Corn Stink Rot : Prevention & Control: Choosing disease-resistant varieties; refraining from continuous cropping; planting at a suitable density to enhance air circulation; promptly removing infected plants. |
| TianGong | During its growth cycle, corn may encounter a myriad of pests and diseases, among which several of the more prevalent include: 1. Northern Corn Leaf Blight: A fungal disease that primarily targets young leaves, while older leaves are severely affected leaf sheaths and husks. [web page] 2. Southern Corn Leaf Blight: Another fungal disease characterized by elliptical, angular, or greyish-green spots that progressively develop into large, irregular, tan-colored areas. [web page] 3. Corn Sheath Blight: A bacterial disease caused by fungi, it primarily impacts the leaf sheath but can also affect leaves, husks, and ears. [web page] These many prevention and control measures for these pests and diseases encompass: 1. Agricultural Management: This involves selecting disease-resistant varieties; practicing rational crop rotation; sowing at appropriate times; and applying fertilizers judiciously. [web page] 2. Biological Control: Harnessing the power of natural enemies such as predatory insects and microbial agents to manage pests and diseases. 3. Chemical Control: Utilizing pesticides for prevention and control, with a keen emphasis on scientific application to minimize environmental and human health risks. |
| Expert | The common diseases affecting corn include Northern Corn Leaf Blight , Southern Corn Leaf Blight , Corn Brown Spot , Corn Rust , Corn Curvularia Leaf Spot , Corn Rootworm , Corn Borer , and Corn Earworm . Prevention and control measures for these diseases involve a combination of cultural, chemical, and biological approaches. The most economical and effective approach is prevention and control using Northern Corn Leaf Blight as an example. 1. Planting disease-resistant varieties : This is the most economical and effective approach. However, when selecting, breeding, and promoting these varieties, it is crucial to consider a rational layout and regular rotation, combined with optimal farming and cultivation practices. 2. Strengthen cultivation management measures by paying attention to the removal of disease sources . 3. Utilize chemical products such as 70% Maneb and 50% Dimecban , but it is imperative to follow the specific application timelines and dosages as outlined in the product manuals or on the Pesticide Information Registration Network. Increasing dosages arbitrarily is strictly prohibited. |
| Our Model | The prevalent diseases affecting corn include Northern Corn Leaf Blight , Southern Corn Leaf Blight , and Corn Rust , while the major pests consist of Corn Borer and Cotton Bollworm . Northern Corn Leaf Blight occurs during the mid-to-late growth stages, while Southern Corn Leaf Blight occurs during the early growth stages. Both diseases can lead to a yield reduction of 10% to 30% in severe cases. Prevention & Control Methods: (1) Select disease-resistant varieties; (2) strengthen field management and the application of phosphorus-potassium fertilizers to enhance plant resistance; (3) promptly remove weeds from the field to reduce pathogen sources; (4) At the initial stages of infection, spray a mixture of 800x diluted 50% Carbendazim Wettable Powder, 600x diluted 75% Chlorthalonil Wettable Powder, and 800x diluted 50% Thiflaphanate-methyl Wettable Powder combined with 600x diluted 75% Chlorthalonil Wettable Powder, following the instructions on the product manual. Note: When spraying, pay attention to the weather conditions occurring during the mid-to-late growth stages; it primarily affects corn leaves and can result in a yield reduction of 20% to 30% in severe cases. Prevention and Control Methods: (1) Select disease-resistant varieties; (2) Corn Rust: Appearing during the mid-to-late growth stages, it primarily targets corn leaves. (4) At the initial stages of infection, spray a mixture of 800x diluted 15% Triadimenol Wettable Powder and 800x diluted 10% Tridemorph Wettable Concentrate and then alternate with 800x diluted 10% Tridemorph Wettable Powder, adhering strictly to the product instructions. 4. Corn Borer: Occurring during the mid-to-late growth stages, it primarily damages corn ears and can cause a yield reduction of 20% to 30% in severe cases. Prevention and Control Methods: (1) Select insect-resistant varieties; (2) Enhance field management, increase phosphorus and potassium fertilizers to boost plant resistance; (3) Regularly remove weeds to minimize insect populations; (4) During the adult oviposition period, use insecticides to spray and vision to kill and eliminate adult insects; (5) During peak larval hatching, apply 1000x diluted 90% Trichlorfon Crystal Solution directly into the hearts of affected plants or spray with 800x diluted 90% Trichlorfon Crystal Solution for control. Cotton Bollworm: Appearing during the mid-to-late growth stages of corn. |

Fig. 1. Responses to crop-specific pest and disease query from top-tier LLMs, search engines, agricultural experts, and our proposed model. GPT-4 [14] and Claude [15] represent state-of-the-art open-domain LLMs, while TianGong [16] showcases results obtained via search engines. Expert denotes the responses given by agricultural experts. Our Model refers to the output generated by the KALLM proposed in this paper.

Some researchers have sought to enhance the domain-specific dialogue understanding and generation capabilities of general large language models by curating domain-relevant dialogue datasets and refining model parameters through supervised fine-tuning (SFT) techniques [17,18]. For instance, in the medical field, Q&A data is often automatically constructed by knowledge graphs or crawled from online medical communities [19–21]. These dialogue datasets constitute an essential resource for transitioning from general LLMs to medical-specific LLMs. Similarly, in the legal domain, bar examination questions and legal consultation transcripts are utilized as corpora for SFT [22, 23]. Although SFT is a potent method to inject parametric knowledge into LLMs, it has several issues. Firstly, most domain dialogue data are generated automatically, lacking standardization and potentially compromising accuracy, particularly in agriculture, where no public dialogue dataset is currently available. Secondly, contemporary LLMs rely on contextual learning (CL) [24] to fine-tune, which employs the preceding context (i.e., the query and previously generated tokens) to predict the next token and adjusts model parameters via a cross-entropy loss. Even if parametric fine-tuning can learn new domain knowledge, the parameterized knowledge encapsulated within LLMs will inevitably result in factual inconsistent responses [25]. Moreover, some studies have revealed that such training methods tend to prioritize the fitting of simple tokens [26–28] and pay insufficient attention to professional tokens that are highly relevant to domain knowledge. It may overlook critical domain knowledge, lead to errors, and pose risks in practical applications.

In order to provide richer contextual information to LLMs, retrieval-augmented generation (RAG) is used to supplement the input query by retrieving relevant evidence from an external knowledge base. This approach helps LLMs facilitate rapid domain adaptation and improve factual consistency [25,29]. However, RAG frameworks exhibit significant limitations. Firstly, most RAG-based models rely on webpages, such as those on Wikipedia, as their knowledge base for evidence retrieval. These sources often suffer from issues of low quality, redundancy, and noise [30–33]. Secondly, the retrievers typically utilize either dense retrieval methods – relying on similarity calculations

through encoding – or sparse retrieval approaches, which depend on exact character matching [28,34–36]. The former can capture deep semantics but requires extensive vector computations. The latter relies on surface-level character matches, which demand fewer computational resources but lack deep semantic understanding. For an agriculture-specific knowledge base, where textual information exhibits a high degree of semantic similarity, dense retrieval struggles to effectively capitalize on its semantic encoding strengths. Meanwhile, sparse retrieval focuses on every character in evidence searching and fails to highlight domain-relevant characters that are usually crucial for evidence localization.

To address the aforementioned challenges, our research primarily delves into three questions: How to construct high-quality agricultural dialogue datasets and knowledge bases? How can language models improve their attention to domain-specific keywords, thus increasing confidence in agricultural decision-making? How to efficiently and precisely pinpoint the corroborative evidence from the knowledge base? Building upon these motivations, we widely collect related agricultural books and national standards. Under the guidance of agromonic experts, we establish guidelines for constructing knowledge bases and annotate 220,000 high-quality agricultural Q&A pairs, named Chinese Crop Advisory Dataset (CCAD). Furthermore, we propose a Knowledge-guided Agriculture LLM (KALLM), which integrates agriculture knowledge at both token-level and sentence-level during the model training and inference stage. To our knowledge, CCAD is the first agricultural dialogue dataset and knowledge base, and KALLM is also the inaugural knowledge-guided LLM for agricultural decision-making. To offer valuable insights for future agricultural LLM, we conduct comprehensive ablation experiments in terms of various data scales, multiple SFT methods, and diverse retrievers. The main contributions of this paper are as follows:

- For four critical agricultural decision-making tasks: crop variety recommendation, cultivation management, crop pest and disease treatment, we extensively gather agricultural books, scientific literature, and national standards as a source of domain knowledge. We then develop a standard guideline for building agricultural Q&A datasets and construct a high-quality Q&A base using a semi-automatic approach.
- To the best of our knowledge, this is the first attempt to propose a knowledge-guided agricultural LLM, which assimilates domain knowledge at both token-level and sentence-level, to enhance the factual consistency and practical feasibility of agricultural decisions. Specifically, we introduce a knowledge-coordinated fine-tuning method and a topic-matched RAG approach that incorporates a self-reflection mechanism for calculating semantic and domain relevance.
- We apply the proposed knowledge-guided framework to strengthen multiple outstanding LLMs, including LLaMa, ChatGLM, Qwen, and Baichuan. Comprehensive experimental results indicate that these optimized LLMs have significant improvements in response fluency, accuracy, authenticity, and domain fidelity. Additionally, we explore the impact of data scale and quality on model performance, providing valuable insights for the construction of agricultural datasets and the development of specific-domain LLM in the future.

2. Related work

There are two prevalent approaches to integrate proprietary and domain-specific data when adapting general LLMs to vertical domains: knowledge injection through implicit parameter supervised fine-tuning (SFT) and retrieval-augment generation (RAG) that leverages explicit external knowledge [18,25]. SFT utilizes supervised training to fit domain-specific Q&A pairs, aligning the predicted token sequences closely with the domain knowledge distribution. To augment prompts

and assist the model in generating factual, consistent responses, RAG retrieves evidence from external knowledge bases to support the given query. The former inherently enhances the model's semantic understanding and generation capabilities for domain-specific issues, while the latter facilitates rapid domain adaptation.

2.1. Domain knowledge assimilation

General LLMs possess remarkable prowess in semantic understanding and language generation. However, owing to the scarcity of domain knowledge during pre-training and SFT, they often generate factually incorrect responses in vertical domain queries [11]. To address the insufficient expertise of LLMs in vertical domains, many studies inject domain knowledge into LLMs through domain-specific dialogue SFT [17–20]. BenCao [20] utilizes medical knowledge graph (KG) triples as information sources to craft dialogue data with general LLMs and fine-tune models like LLaMA-7B and Bloom-7B. HuaTuo [21] constructs dialogues via ChatGPT [6] and gathers real-world conversations to fine-tune LLaMA-13B and Baichuan-7B. DISC-MedLLM [19] trains Baichuan-13B using dialogues built from KG and real medical consultation records. Lawyer [22] fine-tunes LLMs with bar exam questions and legal consultations, while ChatLaw [23] automatically constructs domain Q&A data from forums and legal documents. Although SFT with domain-specific data enables models to generate responses with domain expertise, there are two significant challenges: First, high-quality, specialized training data is scarce. Second, there are no established industry standards for constructing specific-domain Q&A data. These factors currently constitute the core bottlenecks preventing the development of agricultural LLMs.

The researchers also observed that LLMs tended to prioritize reducing general token losses during SFT, focusing more on general token generation while ignoring domain-specific token predictions [26–28]. For instance, concerning the Q&A pair “Query: How to control corn borer chemically? Answer: Use phoxim in larval stage”, existing SFT methods tend to preferentially fit tokens like “Use”, “in”, and even punctuation marks, which evidently deviates from the domain-specific Q&A that prioritizes agricultural keyword generation [37], such as “phoxim” and “larval”. Improving the model’s attention to domain-specific keywords through token-level knowledge assimilation is pivotal for enhancing domain confidence. This is similar to Focus Learning (FL) for open-domain dialogue systems [26] and summarization tasks [27], which adjusts the contribution of tokens to model optimization, thereby increasing the model’s attention to knowledge-related tokens.

In summary, to achieve domain adaptation for general LLMs, it is imperative to imbue the model with high-quality domain-specific knowledge. This necessitates not only annotating high-quality data tailored to agricultural decision-making but also designing an effective SFT strategy to assimilate domain-specific knowledge to the model’s parameters.

2.2. Retrieval-augmented generation

To maintain consistency between responses and knowledge, some researchers have adapted external knowledge sources to correct LLMs’ responses during the inference phase [25]. RAG, which integrates generative models with knowledge retrieval methods, effectively improves the generated context’s relevance, coherence, and factual accuracy [29]. RAG involves several aspects: timing of retrieval (to avoid unnecessary searches) [32], source of retrieved evidence (to provide a high-quality knowledge base) [30,31], retrieval method (to obtain highly relevant evidence) [34–36], and evidence filtering (to exclude noise to further enhance accuracy) [33].

Most RAG methods indiscriminately retrieve evidence and incorporate a fixed number of retrieved passages upon receiving queries regardless of whether retrieval is necessary, which reduces the efficiency of response generation [25,29]. Active RAG [32] proposes to

judge the necessity of retrieval based on the model’s confidence in generating tokens, retrieving evidence only when token probabilities fall below a certain threshold. Similar to the self-check mechanism, self-RAG [33] advocates for on-demand retrieval, employing a reward model to assess whether the query warrants retrieval. On the other hand, current RAG-based models predominantly utilize Wikipedia as their knowledge base [30,31], which not only contains uncleared low-quality noise, but also spans multiple domains, resulting in a lack of focus on specific domains. For retrieval methods, there are two main categories: dense retrieval relies on encoding similarity [34,35], and sparse retrieval centres around character matching [36]. Dense retrieval leverages pre-trained models like BERT [38] and T5 [39] to embed both the input query and evidence from the knowledge base and then select evidence with high semantic similarity to the query through cosine similarity. Sparse retrieval, primarily using Best Matching 25 (BM25) [36] retriever, locates evidence with character similarity to the query. In scenarios with high semantic similarity (e.g., crop disease diagnosis, variety recommendation), dense retrieval does not yield ideal accuracy in evidence acquisition, while sparse retrieval matches all characters in a sentence, failing to highlight domain-specific characters critical for evidence retrieval. For the evidence filtering process, most methods before Self-RAG [33] augment the input context with all top-k retrieved evidence. Self-RAG [33] constructs a dataset to assess evidence relevance via ChatGPT [6] and trains an effectiveness evaluation model for filtering evidence. Some experimental results have proved that the relevance and validity of retrieved evidence are pivotal in improving domain consistency in responses.

To the best of our knowledge, there remains ample room for exploration in agricultural dialogue datasets, knowledge bases, and agricultural decision guidance. Taking into account datasets, knowledge, domain-specific SFT, and external knowledge augmentation, we propose the first high-quality agricultural dialogue dataset and knowledge base, along with the inaugural knowledge-guided agricultural decision LLM, which assimilates knowledge from both the token-level and sentence-level.

3. Method

In this section, we will introduce the proposed knowledge-guided agriculture large language model (KALLM). To incorporate domain knowledge into LLM, we propose a token-level supervised fine-tuning (SFT) (see Section 3.2) and a sentence-level retrieval-augment generation (RAG) (see Section 3.3) to optimize the text decoding and prompt augmentation of LLM, respectively. The KALLM framework is depicted in Fig. 2.

3.1. Problem definition

For a given query sequence $Q = (q_1, q_2, \dots, q_n)$ composed of n tokens and a standard output $Y = (y_1, y_2, \dots, y_m)$, where n, m denotes the number of tokens in Q and Y , respectively, the objective of general LLMs is to maximize the joint distribution of the desired sequence (y_1, y_2, \dots, y_m) . As an autoregressive prediction approach, general LLMs forecast the i th token y_i based on the query Q and the preceding $i - 1$ tokens $y_{<i} = (y_1, \dots, y_{i-1})$, i.e., $P(y_i|Q, y_{<i})$. The objective function is shown in (1).

$$Obj = \max_{\theta} \sum_{(Q,Y)} \sum_{i=1}^m \log P_{\theta}(y_i|Q, y_{<i}) \quad (1)$$

where θ is the trainable parameter. However, due to the lack of focus on knowledge, LLMs trained in this way tend to overlook critical domain information and are more prone to factual errors in their responses. Our proposed knowledge-guided LLM not only generates responses based on the user’s query but also adheres to the constraints imposed by domain knowledge K , ensuring fact-consistent and highly domain-confident

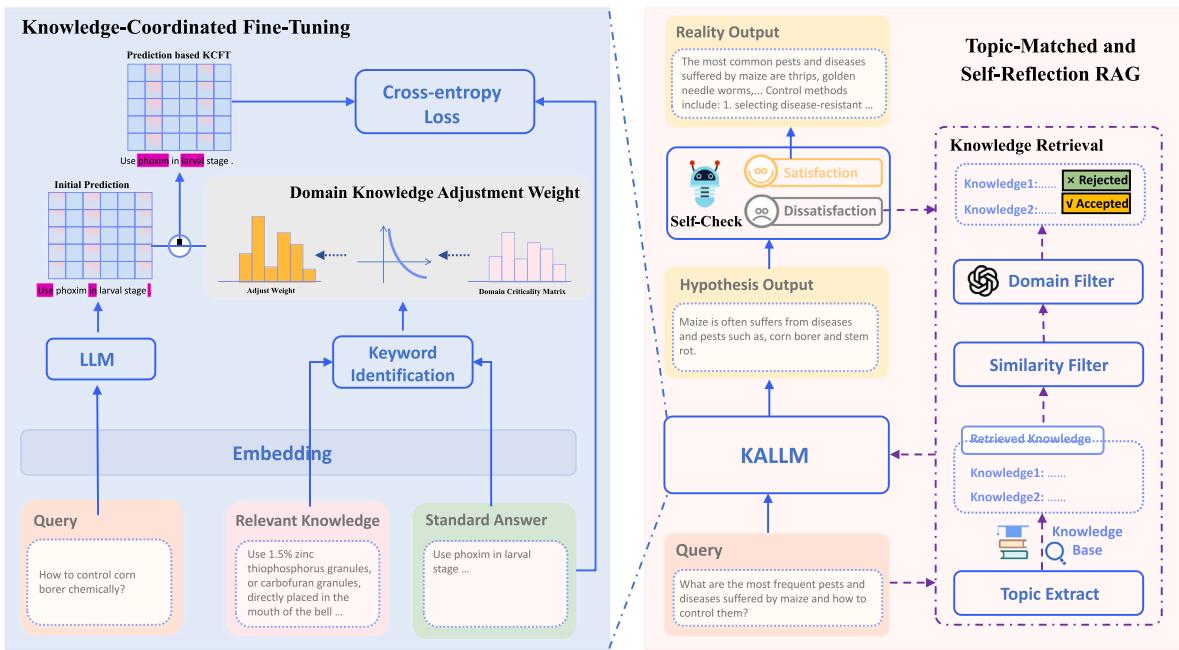


Fig. 2. KALLM consists of knowledge-coordinated fine-tuning and topic-matched RAG. During fine-tuning, we first locate the positions of domain-specific keywords using semantic similarity. By adjusting weights, we amplify the prediction probabilities of these tokens and increase their loss contributions. During inference, our RAG method extracts topic terms from the query to capture supporting evidence in the knowledge base and uses domain evaluations and semantic relevance to achieve dual-step evidence filtering.

decision guidance. As a result, the objective function is optimized to (2).

$$Obj = \max_{\theta} \sum_{(Q, Y)} \sum_{i=1}^m \log P_{\theta}(y_i | Q, y_{<i}, K) \quad (2)$$

3.2. Knowledge-coordinated fine-tuning

SFT is a pivotal method for integrating domain knowledge into general LLMs. Given a query Q , LLMs produce a predicted answer sequence $A = (a_1, a_2, \dots, a_m)$, where $a_i \in \mathbb{R}^L$ represents the token distribution at the i th position of the output sequence and L is the vocabulary size. Traditional fine-tuning methods, such as LoRA [18] and full FT [40], adjust model parameters by calculating the cross-entropy loss between the predicted answer A and the standard response Y , as shown in formula (3). Although the fine-tuning method partly integrates domain knowledge into model parameters, some studies have revealed that it tends to fit general tokens (i.e., preferentially reducing the autoregressive loss corresponding to general tokens) over domain-specific keywords [26,27], resulting in lacking domain salience and missing crucial domain-specific information.

$$L_O = -\frac{1}{m} \sum_{i=1}^m y_i \cdot \log P(a_i | Q, a_{<i}) \quad (3)$$

To amplify the LLM's attention towards domain-specific keywords and elevate its domain confidence, we propose a knowledge-coordinated fine-tuning (KCFT) method, which scales the contribution of predicted tokens to the final loss based on their relevance to domain knowledge.

The KCFT method focuses on two critical aspects: (1) how to identify domain-relevant keywords during the fine-tuning training process, and (2) how to utilize these domain-specific keywords to achieve token-level fine-tuning.

3.2.1. Keyword identification

To adjust the contribution of different tokens to the autoregressive loss and pay more attention to agriculture-specific tokens, we first need to define domain-specific tokens.

For a Q&A pair (Q_i, Y_i) , we project each token y_{ij} in the standard response sequence $Y_i = (y_{i1}, y_{i2}, \dots, y_{im})$ and the corresponding knowledge K_i associated with the pair into a shared semantic space, where y_{ij} represents the j th token in Y_i . In this paper, we employ BERT [38] to embed y_{ij} and K_i and utilize the vector corresponding to [CLS] to represent their semantic encodings, as shown in (4) and (5). Subsequently, we evaluate the domain expertise of each token by calculating the semantic distance between y_{ij} and K_i , as expressed in (6).

$$\hat{y}_{ij} = Bert(y_{ij}) \quad (4)$$

$$\hat{K}_i = Bert(K_i) \quad (5)$$

$$S_{ij}(y_{ij}, K_i) = \frac{\hat{y}_{ij} \cdot \hat{K}_i}{\|\hat{y}_{ij}\| \cdot \|\hat{K}_i\|} \quad (6)$$

where \hat{y}_{ij} and \hat{K}_i represent the semantic embeddings of token y_{ij} and knowledge K_i , respectively, and $S_{ij}(y_{ij}, K_i)$ denotes the semantic cosine similarity between them. Based on the above computations, we obtain the domain criticality matrix $S_i \in \mathbb{R}^m$ corresponding to the standard response sequence Y_i in the training dataset. The matrix measures which tokens are more crucial in the context of the knowledge relevant to the current domain-specific query. If a token y_{ij} in Y_i is highly similar to the corresponding domain knowledge K_i , the value of S_{ij} will be larger, indicating that the token possesses higher domain specificity.

3.2.2. Adjust weight and focused cross-entropy loss

To realize token-level knowledge guidance according to captured domain keywords, we draw inspiration from Focus Learning [26,27] and adapt the domain criticality matrix S_i to amplify the output probabilities of the predicted token distribution at positions corresponding to domain-related tokens, thereby magnifying the impact of these tokens' predictions on the loss function. For a query Q_i , the adjustment weight for the prediction probability of the j th token in the generated sequence is calculated as follows in (7):

$$w_{ij} = -\ln(1 - S_{ij}(y_{ij}, K_i) + \lambda) + 1 \quad (7)$$

where $\lambda \in (0, e-2)$ serves as a non-zero constraint factor. After acquiring the adjustment weights of token-level domain knowledge, we scale

the original predicted distribution A and compute the probability of each token using the new logit. Given the model's initial predictions $A_i = (a_{i1}, a_{i2}, \dots, a_{im})$, the output probability of tokens consistent with domain knowledge is expressed as (8):

$$P_w(\beta_{ij} | Q_i, K_i, \beta_{i,<j}) = \text{softmax}(w_{ij} \cdot a_{ij}) \quad (8)$$

$\beta_{ij} \in \mathbb{R}^L$ represents the prediction probability of the j th token by the model, with L denoting the vocabulary size of the model, and $\beta_{i,<j}$ represents the tokens generated by the model prior to the j th token for query Q_i . $P_w(\beta_{ij} | Q_i, K_i, \beta_{i,<j})$ indicates the probability of the next token prediction by the model, adjusted by agricultural knowledge. The cross-entropy optimization after scaling with domain knowledge is rewritten from (3) to (9):

$$\mathcal{L}_{KC} = -\frac{1}{m} \sum_{i=1}^m y_{ij} \cdot \log P(\beta_{ij} | Q_i, K_i, \beta_{i,<j}) \quad (9)$$

This formula amplifies the contribution of tokens, which are highly correlated with domain knowledge, to the loss calculation, thereby enhancing the model's focus on domain-specific keywords and achieving supervised fine-tuning with token-level knowledge collaboration.

3.3. Topic-matched and self-reflection RAG

Knowledge-coordinated supervised fine-tuning effectively internalizes domain knowledge into its parameters. However, it is challenging to efficiently utilize the domain knowledge stored within the parameters due to the inherent predictive mechanism of autoregressive models. The prompt is a considered tool for accessing internal model knowledge [25]. To activate internal knowledge within the model and expedite domain adaptation, we introduce an RAG based on an agricultural external knowledge base $K = \{K_i\}_{i=1}^d$ to augment the post-knowledge-coordinated fine-tuned LLM M^* . This transforms the LLM's prediction from $P_M^*(Y|Q)$ to $P_M^*(Y|Q, K)$.

To ensure the efficacy of retrieval and maintain domain consistency, precision, and accuracy in evidence acquisition, we propose a Topic-matched and Self-reflection RAG approach (abbreviated as TSRAG). TSRAG primarily involves three aspects: 1. When to retrieve: To curtail redundant retrievals, we implement a model self-check mechanism. It evaluates the congruence between the query and hypothetical response; output ensues upon an “accept” decision, otherwise, retrieval commences (see Section 3.3.1). 2. How to retrieve: In domain tasks with high semantic similarity, dense retrieval loses its edge, yet sparse retrieval requires matching each token individually, failing to emphasize the significance of domain-specific terms typically crucial for evidence pinpointing. Therefore, we devise a topic-matched retrieval method, extracting topic words from the query instead of using all tokens for evidence retrieval from the knowledge base, ensuring the acquisition of more relevant evidence (see Section 3.3.2). 3. Filtering retrieval noise: To enhance the accuracy of evidence support, we establish a self-reflection mechanism. This mechanism filters retrieval evidence based on semantic and domain relevance (see Section 3.3.3).

3.3.1. Hypothetical response and self-check mechanism

To mitigate unwarranted knowledge base accesses, we first conduct self-checking during the inference stage on the query Q and the preliminary sequence generated by M^* , denoted as hypothetical response $H = (h_1, h_2, \dots, h_m)$. Employing an evaluation model C^1 , we solicit feedback on H 's rationality via $C^1(Q, H)$. The feedback comprises decisions: “Accept”, “Reject”, or “Retrieve”. When the result is “Accept”, the output is finalized; otherwise, regeneration or retrieval is initiated.

3.3.2. Evidence retrieval based on topic matching

During the evaluation of retrieval algorithms, we find that using an encoder to retrieve evidence by calculating embedding similarity is challenging in cases where Q&A closely resembles associated knowledge (see analysis in 4.3). Additionally, sparse matching is sensitive to all tokens, failing to highlight the pivotal role of domain-specific terms within queries. Therefore, we devise an evidence retriever \mathcal{R} centred on topic matching.

Initially, we employ a topic term extraction model \mathcal{T} to distill a list of domain-specific topic words T from the query, as shown in (10).

$$T = (t_1, t_2, \dots, t_k) = \mathcal{T}(Q) \quad (10)$$

where each topic term appears in Q . Subsequently, we initiate a fuzzy matching process against the knowledge base items for each topic term. The union of all retrieved knowledge items forms the initial set of external evidence:

$$D_{t_i} = \mathcal{R}(t_i) \quad (11)$$

$$D_Q = \mathcal{R}(Q) = D_{t_1} \wedge D_{t_2} \wedge \dots \wedge D_{t_k} \quad (12)$$

where D_{t_i} represents the evidence retrieved using the topic term t_i , and D_Q denotes the collection of all evidence from the knowledge base supporting the query Q .

3.3.3. Dual-step filtering mechanism

To ensure the semantic coherence and domain consistency of the external knowledge used for prompt augmentation, we evaluate the retrieved knowledge items based on semantic similarity and domain relevance.

Step 1: Semantic Similarity. We semantically encode each piece of evidence in D_Q alongside the query Q and compute their semantic similarity as (13). By this, we rank the evidence and retain the top-k most similar evidence D'_Q for the subsequent filtering step.

$$S(d_i^r, Q) = \frac{d_i^r \cdot Q}{\|d_i^r\| \cdot \|Q\|} \quad (13)$$

where $d_i \in D_Q$, $d_i^r = DPR(d_i)$ and $Q = DPR(Q)$ denote the semantic representation of evidence d_i and Q via Dense Passage Retrieval (DPR) [34], respectively. Like BERT, DPR is a vector-based semantic encoding method built upon T5 [39]. We train it on a portion of our training set through contrastive learning [41] to enhance its domain encoding capabilities.

Step 2: Domain Relevance. For each $d_i \in D'_Q$, we deploy the assessment model C^2 to evaluate its speciality to the domain, i.e., $C^2(d_i, Q)$. The final filtered and precisely matched set of evidence is denoted as D_Q^* . Due to ChatGPT's exceptional prompt comprehension and its intrinsic vast general knowledge, we adopt it as both the evaluation model and the topic term extractor. Fig. 3 illustrates the prompt employed for topic term extraction.

Therefore, the knowledge-guided domain decision model ultimately generates responses as follows:

$$Y = M^*([Q, D_Q^*]) \quad (14)$$

4. Experiment

4.1. Experimental setup

We implement KALLM based on the PyTorch framework and conduct training and inference on one NVIDIA A800 80G GPU. The meanings and specific settings of each hyper-parameter involved in the model are detailed in Table 1. Since KALLM is used to enhance the ability of multiple general models to assimilate agricultural knowledge, we follow all settings from the original papers corresponding to each general LLM.

Table 1

Hyper-parameter settings. The meanings and specific settings of each hyper-parameter.

| Hyper-parameter | Meaning | Setting |
|-----------------------|---------------------------------------|--------------------------------|
| batch_size | Batch size of training | 4 |
| gradient_accumulation | Gradient accumulation steps | 4 |
| lr_scheduler_type | Learning rate scheduler | Cosine learning rate scheduler |
| learning_rate | Initial learning rate | 5.00E-05 |
| num_train_epochs | Number of training epochs | 3 |
| Temperature | Diversity control in model generation | 0.7 |

Table 2

The parameter sizes, the number of tokens utilized during pre-training, and the layers of Transformer of baseline backbones.

| Base Backbones | Size of Model Parameters | Number of Training Tokens | Transformer Blocks |
|--------------------|--------------------------|---------------------------|--------------------|
| Baichuan2-7B-Chat | 7B | 2.6T | 32 |
| Baichuan2-13B-Chat | 13B | 2.6T | 40 |
| ChatGLM3-6B | 6B | 1.4T | 28 |
| LLaMA2-7B-Chat | 7B | 2T | 32 |
| LLaMA2-13B-Chat | 13B | 2T | 40 |
| Qwen1.5-7B-Chat | 7B | 3T | 32 |
| Qwen1.5-14B-Chat | 14B | 3T | 40 |

4.2. Baselines

To verify the performance of our method, we conduct comparative experiments across multiple backbones: Baichuan-7B [7], Baichuan-13B [7], ChatGLM3-6B [42], Llama2-7B [8], Llama2-13B [8], Qwen1.5-7B [9], Qwen1.5-14B [9], and HuaTuo [21], all of which apply the Transformer architecture to autoregressively predict answers for given queries. Table 2 outlines the parameter sizes, the number of tokens utilized during pre-training, and the layers of Transformer for each backbone above. To verify the lack of architectural decision-making in general LLMs and the insufficient attention to domain knowledge in existing fine-tuning methods, we compare our approach with GPT-4, RAG-based GPT-4, as well as the traditional fine-tuning and retrieval-augment generation pipeline.

4.3. Dataset

Under the advice of agricultural experts, we initially identify four prominent agricultural decision-making tasks: crop variety recommendation, cultivation management, crop pest and disease diagnosis and prevention. To construct the dialogue dataset and knowledge base, we widely collect related professional books, scientific literature, and national standards from various high-quality repositories (like government agencies and scientific knowledge databases) as the foundational information sources. To ensure the precision and domain specificity of the annotated Q&A pairs, we formulate data construction standards under the guidance of agronomists. The annotation guidelines specify the essential elements, data source format, and annotation format for Q&A pairs in each task. Table 3 lists the mandatory factors for different tasks. Q&A pairs are annotated through “forward” and “reverse” to ensure diversity. Appendix B provides related examples.

To avoid bias in the data construction, we hire dozens of data annotators with agricultural backgrounds to annotate the Q&A pairs in a cross-validation manner. Ultimately, we establish a high-quality agricultural dataset consisting of 220,000 dialogues and a corresponding knowledge base of more than 15 million tokens. Each Q&A pair is annotated and reviewed by at least two people, with automated consistency evaluation conducted through ChatGPT. Only those with an inter-annotator agreement of 85% or higher are retained. Each item

Domain Topic Extraction Prompt

Extract the domain topics from the following query. Ensure that the extracted topics can be used as key strings for searching for evidence in the knowledge base. Use commas to separate multiple topics. Here are some examples of extracted topics from queries:

Query 1: What corn varieties are recommended for planting in the mid-late maturity spring corn type regions in Liaoning Province with an active accumulated temperature of $\geq 10^{\circ}\text{C}$ over 2800°C ?

Extraction Result 1: [Liaoning Province $\geq 10^{\circ}\text{C}$ active accumulated temperature over 2800°C , mid-late maturity spring corn type regions, recommended corn varieties]

Query 2: The corn is now tasseling, but some have swollen lumps on the stalk and ear, initially white or light green, later with black stripes, and the serious lumps have powdery substances. What disease does the corn have?

Extraction Result 2: [corn tasseling stage, stalk, and ear swellings, white or light green swellings, black stripes, powdery substances, corn disease diagnosis]

Query 3: How to control wheat stem rust?

Extraction Result 3: [wheat stem rust, wheat stem rust control]

Query: {insert query}

Extraction Result: {domain topics in the query}

Fig. 3. Domain topic extraction prompt.

of annotated content is shown in Appendix B. The number of Q&A pairs, the total token count, the average tokens per entry, and the distribution of QA pairs across tasks are detailed in Table 4. An evaluation of the dataset’s quality is provided in Section 4.5.

During the knowledge-coordinated fine-tuning (KCFT) and inference phase with topic-matched and self-reflection retrieval-augment generation (TSRAG), we randomly allocate 80% of each task’s data for training and reserve 20% for testing. More data samples can be found in Appendix B.

4.4. Evaluation metrics

We use standard text generation evaluation metrics such as BLEU, ROUGE, and BertScore to assess the fluency, accuracy, and domain confidence of the generated responses. BLEU and ROUGE evaluate the fluency and accuracy of the generated responses by comparing them to the reference answers using N-gram splitting and calculating precision and recall. BertScore leverages BERT embeddings to assess the domain confidence of the generated responses by computing the cosine similarity between the model-generated outputs and the reference answers. The formulas for these evaluation metrics are presented in Appendix A.

4.5. Data quality analysis

4.5.1. Comparison of dataset quality

To validate that datasets extracted from books and manually constructed are more effective at assimilating implicit domain knowledge into model parameters, we build another batch of Q&A data using a fully automated method for comparison. We crawl a substantial amount of agricultural Q&A pairs from relevant forums¹ and use ChatGPT to automatically construct the Q&A dataset from agricultural books. After deduplication and other data-cleaning processes, we obtain 110,000 Q&A pairs.

¹ <http://zjzx.cnki.net/BBS/BBSList?code=101001>

Table 3

Factors involved in agricultural decision-making data construction.

| Task | Factor | Information Source |
|----------------------------------|---|---|
| Crop Variety Recommendation | Local Environmental Information (including but not limited to time, planting area, accumulated temperature, soil information, weather information, mechanization level, etc.); The Types and Varieties of Crops | "National Catalog of Promoted Elite Crop Varieties" |
| Cultivation Management | Crop Varieties ; Land Parcel Information ; Weather Conditions ; Mechanization Level ; Standardized Planting Procedures for Each Growth Stage | "Cultivation of Rice in Northeast China and Comprehensive Plant Protection Technology Manual" |
| Crop Pest and Disease Diagnosis | The Type of Crop ; Description of the Symptom s; a Preliminary Diagnosis Result | "Crop Diseases and Insect Pests in China" |
| Crop Pest and Disease Prevention | The Diagnosis of Pests, Diseases, and Weeds ; Basic Prevention and Control Measures Along With Medication Recommendations (recommendations Should Include Precise Medication Types, Dosage, Water-soluble Ratio, Precautions, and Other Relevant Details) | "Crop Diseases and Insect Pests in China" |

Table 4

Data statistics. The number of Q&A pairs, the total token count, the average tokens per entry, and the distribution of QA pairs across tasks. "#Avg. Tokens/Q&A Pair" signifies the average number of tokens contained within each Q&A pair, "#Avg. Tokens/Knowledge Item" represents the average number of tokens in each knowledge item.

| Data Type | Data Statistics |
|--|-----------------|
| # Q&A Pairs | 227,542 |
| # Tokens in Q&A Pairs | 34,866,682 |
| #Avg. Tokens/Q&A Pair | 153 |
| # Tokens in Knowledge Base | 15,812,576 |
| # Knowledge Items | 113,125 |
| # Avg. Tokens/Knowledge Item | 140 |
| # Q&A Pairs for crop variety recommendation | 113,854 |
| # Q&A Pairs for cultivation management | 111,374 |
| # Q&A Pairs for crop pest and disease diagnosis | 574 |
| # Q&A Pairs for crop pest and disease prevention | 1,740 |

Based on the traditional partial fine-tuning approach (LoRA), we use the manually annotated and automatically constructed datasets to fine-tune general LLMs mentioned in Section 4.2, respectively. The fine-tuning dataset consists of 100,000 Q&A pairs sampled from each of the two datasets, and the overall test dataset consists of 20,000 Q&A pairs (sampled 10,000 pairs from each of the two datasets). Note that the testing dataset does not overlap with the training data. The performance of the fine-tuned models is shown in Table 5.

It is evident that models trained on manually annotated datasets outperform those trained on web-scraped and automatically constructed datasets in terms of fluency and domain confidence, with average metric improvements of 0.860, 2.113, 1.259, 0.884, 0.644, 1.225, 0.681, 0.090, 0.680, and 8.8%, respectively. This is no doubt that the quality of data influences the direction of parameter optimization. Data extracted from books and constructed by humans contain more domain-specific information and less redundant noise than web-scraped and automatically constructed data. Thus, high-quality data will guide the model in obtaining higher domain confidence. Examples of responses from fine-tuned models on both datasets, alongside agricultural experts' evaluations, can be found in Appendix C.

4.5.2. Human evaluation on data quality

To further analyse the data quality from a professional perspective, we randomly collect 1000 dialogue samples from both the manually annotated and automatically constructed datasets. Three agricultural experts are invited to conduct a human evaluation of the quality of these data. The review focuses on several aspects: fluency, professionalism, informativeness, operability, and overall quality. Fluency: Assesses whether the data and responses are coherent, natural, and easy to understand. Professionalism: Evaluates whether the text meets domain-specific standards and effectively conveys professional knowledge or information. Informativeness: Assesses whether the text provides comprehensive, in-depth information that effectively answers the question

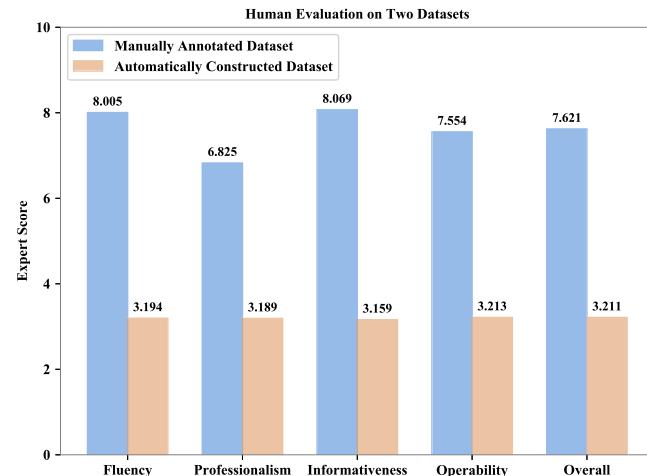


Fig. 4. Expert evaluation on data quality.

or meets the need. Operability: Refers to whether the text offers clear, actionable advice or steps for the user. Overall quality: A comprehensive assessment. Each aspect is rated on a scale of 1 to 10, with 10 being the highest. The Fleiss' Kappa [43] among three agricultural experts is 0.89.

As shown in Fig. 4, the evaluation results indicate that the manually annotated dataset provides more detailed responses to domain-specific queries and has been endorsed by experts across multiple dimensions.

4.5.3. The impact of data scale

Research indicates that the larger the scale of domain-specific data utilized for fine-tuning, the better the model's performance [44]. To explore the relationship between data scale and model efficacy, we incrementally adjust the proportion of the fine-tuning dataset by sampling 50%, 60%, 70%, 80%, and 90% of the training samples. We adopt Qwen1.5-7B as the foundational model. The remaining 10% of the dataset serves as the test set to evaluate the model's performance in the agricultural domain after fine-tuning with different amounts of data. The results are portrayed in Fig. 5.

The comparison results show that the model's performance in the agricultural domain generally improves as the scale of training data increases. Although there is a huge gap in scale between the 220,000 agricultural Q&A pairs and the general dataset measured in TB units, the fluency evaluation represented by the BLUE series is still significantly improved through the fine-tuning of the model. It illustrates that our dataset can effectively promote the understanding and generation of the model for the agricultural context.

Table 5

Comparison of model performance using the manually annotated dataset (Manual) and the automatically constructed dataset (Automatic). GOOGLE_BLEU, BLEU, and ROUGE are presented as natural scores; BertScore is expressed as a percentage. The optimal values under each evaluation criterion are highlighted in bold and green. “Average” represents the mean value achieved by each backbone under different evaluation criteria. The numeric value in the top right corner of “Average(Manual)” indicates its improvement over “Average(Automatic)”. “***”denotes statistical significant with $p < 0.01$.

| Backbone | Data | BLEU** | | | | | ROUGE** | | | BertScore** | |
|--------------|-----------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|------------------------------|
| | | BLEU_1** | BLEU_2** | BLEU_3** | BLEU_4** | Mean_BLEU** | ROUGE_1** | ROUGE_2** | ROUGE_L** | | |
| Baichuan-7b | Manual | 2.097 | 4.243 | 2.323 | 1.598 | 1.154 | 2.329 | 1.126 | 0.074 | 1.126 | 0.679 |
| | Automatic | 1.422 | 2.611 | 1.243 | 0.824 | 0.589 | 1.317 | 0.191 | 0.000 | 0.191 | 0.581 |
| Baichuan-13b | Manual | 2.073 | 4.444 | 2.481 | 1.730 | 1.247 | 2.475 | 0.909 | 0.087 | 0.902 | 0.689 |
| | Automatic | 1.840 | 2.967 | 1.414 | 0.949 | 0.678 | 1.502 | 0.060 | 0.000 | 0.060 | 0.585 |
| ChatGLM3-6B | Manual | 1.610 | 4.060 | 2.297 | 1.605 | 1.173 | 2.284 | 1.240 | 0.095 | 1.240 | 0.681 |
| | Automatic | 0.947 | 1.452 | 0.749 | 0.509 | 0.364 | 0.769 | 0.218 | 0.000 | 0.218 | 0.565 |
| Llama2-7B | Manual | 1.958 | 3.271 | 1.787 | 1.245 | 0.904 | 1.802 | 0.529 | 0.079 | 0.529 | 0.651 |
| | Automatic | 0.699 | 0.783 | 0.405 | 0.280 | 0.203 | 0.418 | 0.087 | 0.000 | 0.087 | 0.577 |
| Llama2-13B | Manual | 2.259 | 3.706 | 2.003 | 1.351 | 0.977 | 2.009 | 0.870 | 0.101 | 0.868 | 0.663 |
| | Automatic | 0.791 | 1.339 | 0.707 | 0.490 | 0.363 | 0.724 | 0.164 | 0.000 | 0.164 | 0.581 |
| Qwen1.5-7b | Manual | 2.271 | 3.765 | 2.066 | 1.430 | 1.041 | 2.076 | 0.755 | 0.114 | 0.755 | 0.679 |
| | Automatic | 1.380 | 1.621 | 0.846 | 0.580 | 0.419 | 0.866 | 0.281 | 0.000 | 0.281 | 0.593 |
| Qwen1.5-14B | Manual | 2.583 | 4.453 | 2.455 | 1.703 | 1.230 | 2.460 | 0.823 | 0.078 | 0.821 | 0.687 |
| | Automatic | 1.749 | 2.379 | 1.233 | 0.839 | 0.600 | 1.262 | 0.481 | 0.000 | 0.481 | 0.629 |
| Average | Manual | 2.121^{10.860} | 3.992^{12.113} | 2.202^{11.299} | 1.523^{10.884} | 1.103^{10.644} | 2.205^{11.225} | 0.893^{10.681} | 0.090^{10.090} | 0.891^{10.680} | 0.676^{18.8%} |
| | Automatic | 1.261 | 1.879 | 0.942 | 0.639 | 0.459 | 0.980 | 0.212 | 0.000 | 0.212 | 0.587 |

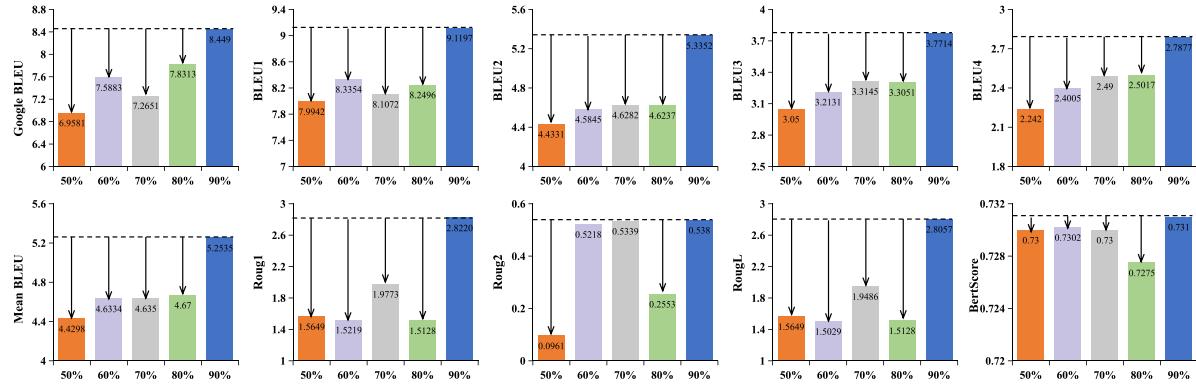


Fig. 5. Comparison of model performance after training with datasets of different scales.

4.6. Model comparison results

4.6.1. Automatic evaluation

We utilize 80% of the manually annotated dataset for training and reserve the remaining 20% for testing. The comprehensive experimental results are encapsulated in [Table 6](#). “Base” refers to the unmodified baseline LLMs, and “Norm” denotes the LLMs trained through traditional fine-tuning (LoRA) and RAG (BM25 retriever) pipeline. “Ours” represents that LLMs achieve knowledge assimilation under our proposed framework that integrates KCFT with TSRAG. To affirm the pivotal role of high-quality domain-specific data in enhancing model performance, we compare our method with GPT-4, which excels in general dialogue tasks. “GPT-4 (TSRAG)” adapts GPT4 with our proposed TSRAG.

It can be seen from [Table 6](#) that our proposed method significantly enhances the domain-specific decision-making capabilities of general LLMs. Compared to GPT-4, our KALLM approaches based on different general LLMs show marked improvements across ten evaluation metrics, with average increases of 29.3679, 21.9460, 17.7579, 15.0816, 13.0765, 16.9655, 20.9298, 9.9175, 20.9249, 26.66%, respectively. Against the GPT-4 with TSRAG, the average effect of the general LLMs optimized by KALLM achieves improvements of 26.0962, 14.8148, 13.3210, 11.6114, 10.2080, 12.4888, 16.2537, 8.3140, 16.2488, 12.63%. Compared to the baseline general LLMs and the traditional pipeline, our knowledge-assimilative method still achieves optimal results on most metrics, validating the effectiveness of the KCFT and TSRAG.

The comparison between GPT-4 and GPT-4 with TSRAG indicates that merely augmenting external knowledge in general LLMs does not significantly improve their performance in vertical domains. We attribute this to the general LLMs’ insufficient understanding of domain-specific dialogues. Even if a relevant knowledge base is provided as a prompt augmentation, their lack of comprehension still leads to suboptimal performance. This analysis is further elaborated in [Section 4.7](#). Compared to GPT-3.5 and Claude, GPT-4 demonstrates better foundational and knowledge understanding, likely due to the incorporation of more knowledge during its training.

Adapting general LLMs to the agricultural domain via our method, KALLM will help these models generate responses that are richer in domain knowledge, stronger factual consistency, and more accurate in information.

In addition, we compare the specific performance of different models across four tasks, as detailed in [Table F.1](#). We find that tasks with larger datasets (e.g., cultivation management) show better performance after training. Preliminary analysis suggests that the number of Q&A pairs involved in training impacts the model’s knowledge injection. The more Q&A pairs related to the domain knowledge, the more knowledge can be infused [44]. Therefore, it is crucial to ensure sufficient data coverage when making task-specific decisions. From the classification comparison, we also observe that, compared to cultivation management, the other three tasks are more challenging. This is because agricultural process flows are relatively fixed, and the corresponding knowledge is easier to learn. In contrast, variety recommendation and pest and disease diagnosis typically involve predicting specific variety

Table 6

Comprehensive model comparison. The best scores for each evaluation metric are highlighted in bold and green. The second-best scores are highlighted in yellow. The numeric value in the top right corner of “Average KALLM (Ours)” signifies its improvement over “GPT4 (Base)”.

| Backbone | Model | GOOGLE BLEU** | BLEU** | | | | | ROUGE** | | | BertScore** |
|--------------|-------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-----------------------------------|-------------------------------------|-----------------------------------|
| | | | BLEU_1** | BLEU_2** | BLEU_3** | BLEU_4** | Mean_BLEU** | ROUGE_1** | ROUGE_2** | ROUGE_J** | |
| GPT3.5 | Base | 0.1574 | 0.0241 | 0.0112 | 0.0071 | 0.0048 | 0.0118 | 0.1843 | 0.0000 | 0.1843 | 0.4882 |
| | TSRAG | 0.5781 | 4.7600 | 2.5848 | 1.8185 | 1.3526 | 2.6290 | 2.6178 | 0.8704 | 2.6178 | 0.5539 |
| GPT4 | Base | 0.1530 | 0.1321 | 0.0892 | 0.0659 | 0.0506 | 0.0844 | 0.2854 | 0.0000 | 0.2854 | 0.4953 |
| | TSRAG | 3.4247 | 7.2633 | 4.5261 | 3.5361 | 2.9191 | 4.5611 | 4.9615 | 1.6036 | 4.9615 | 0.6356 |
| Claude3.5 | Base | 0.1498 | 0.1389 | 0.0796 | 0.0542 | 0.0382 | 0.0777 | 0.2468 | 0.0000 | 0.2468 | 0.5119 |
| | TSRAG | 0.8807 | 2.3955 | 1.1750 | 0.7423 | 0.5037 | 1.2041 | 1.7705 | 0.2793 | 1.7705 | 0.5900 |
| Baichuan-7b | Base | 0.6978 | 0.6869 | 0.3373 | 0.2232 | 0.1584 | 0.3515 | 0.3093 | 0.0000 | 0.3093 | 0.5211 |
| | Norm | 39.2138 | 24.7590 | 20.5666 | 17.9511 | 15.9179 | 19.7987 | 22.4355 | 12.7506 | 22.4355 | 0.7694 |
| KALLM(Ours) | | 42.8386 | 26.2967 | 22.2469 | 19.6876 | 17.6495 | 21.4702 | 23.9533 | 13.8807 | 23.9533 | 0.7706 |
| Baichuan-13b | Base | 0.8785 | 1.6810 | 1.0381 | 0.7260 | 0.5234 | 0.9921 | 1.0127 | 0.0000 | 1.0127 | 0.5537 |
| | Norm | 21.3198 | 14.6662 | 9.3974 | 6.8600 | 5.2095 | 9.0333 | 9.8600 | 1.9986 | 9.8275 | 0.7533 |
| KALLM(Ours) | | 46.6057 | 23.5236 | 20.4608 | 18.5148 | 16.9417 | 19.8602 | 23.0088 | 15.3180 | 23.0088 | 0.7641 |
| ChatGLM3-6B | Base | 0.2083 | 0.0448 | 0.0239 | 0.0160 | 0.0112 | 0.0240 | 0.2743 | 0.0000 | 0.2743 | 0.4916 |
| | Norm | 11.3760 | 12.6284 | 9.3974 | 7.2497 | 5.7355 | 8.7527 | 12.9296 | 4.1262 | 12.9296 | 0.7002 |
| KALLM(Ours) | | 19.0147 | 13.6748 | 10.9761 | 9.3027 | 8.0220 | 10.4939 | 12.8622 | 6.3277 | 12.8622 | 0.7359 |
| Llama2-7B | Base | 0.0212 | 0.0174 | 0.0047 | 0.0025 | 0.0015 | 0.0065 | 0.0093 | 0.0000 | 0.0093 | 0.4147 |
| | Norm | 12.9479 | 12.4864 | 9.1893 | 7.3944 | 6.1071 | 8.7943 | 9.4336 | 4.2052 | 9.4336 | 0.7301 |
| KALLM(Ours) | | 19.7726 | 11.6220 | 9.1452 | 7.8838 | 7.0323 | 8.9208 | 9.2546 | 5.7006 | 9.2211 | 0.7264 |
| Llama2-13B | Base | 0.0201 | 0.0118 | 0.0042 | 0.0024 | 0.0016 | 0.0050 | 0.0147 | 0.0000 | 0.0147 | 0.4152 |
| | Norm | 16.2231 | 13.2787 | 9.9933 | 8.3035 | 7.1185 | 9.6735 | 10.0424 | 5.0443 | 10.0424 | 0.7339 |
| KALLM(Ours) | | 22.8906 | 17.8026 | 13.8248 | 11.6491 | 10.0199 | 13.3241 | 14.5396 | 7.0794 | 14.5396 | 0.7540 |
| Qwen1.5-7b | Base | 0.1440 | 0.0334 | 0.0187 | 0.0128 | 0.0091 | 0.0185 | 0.2814 | 0.0000 | 0.2814 | 0.4962 |
| | Norm | 21.8145 | 15.8999 | 13.1363 | 11.5750 | 10.4394 | 12.7626 | 13.9439 | 8.7080 | 13.9439 | 0.7405 |
| KALLM(Ours) | | 50.1202 | 28.7252 | 24.5268 | 21.6901 | 19.5137 | 23.6140 | 27.7619 | 15.9170 | 27.7619 | 0.7799 |
| Qwen1.5-14B | Base | 0.2230 | 0.0436 | 0.0230 | 0.0154 | 0.0108 | 0.0232 | 0.2620 | 0.0000 | 0.2620 | 0.5048 |
| | Norm | 35.9919 | 19.8655 | 16.9220 | 14.9485 | 13.3445 | 16.2701 | 18.6985 | 11.5615 | 18.6985 | 0.7259 |
| KALLM(Ours) | | 34.6427 | 22.0391 | 18.0749 | 15.7333 | 13.9787 | 17.4565 | 19.7939 | 11.5174 | 19.7939 | 0.7741 |
| HuaTuo | Base | 0.0951 | 0.7030 | 0.3886 | 0.2607 | 0.1810 | 0.3833 | 0.3083 | 0.0000 | 0.3083 | 0.4948 |
| | Norm | 5.0714 | 38.4849 | 27.9786 | 20.5146 | 15.1217 | 25.5250 | 33.0007 | 0.7500 | 33.0007 | 0.7361 |
| KALLM(Ours) | | 0.2821 | 32.9406 | 23.5210 | 16.7186 | 11.8590 | 21.2598 | 38.5476 | 3.5994 | 38.5419 | 0.7905 |
| Average | Base | 0.2860 | 0.4027 | 0.2298 | 0.1574 | 0.1121 | 0.2255 | 0.3090 | 0.0000 | 0.3090 | 0.4865 |
| | Norm | 20.4948 | 19.0086 | 14.5726 | 11.8496 | 9.8743 | 13.8263 | 16.2930 | 6.1430 | 16.2890 | 0.7362 |
| KALLM(Ours) | | 29.5209 ^[29.3679] | 22.0781 ^[21.9460] | 17.8471 ^[17.7579] | 15.1475 ^[15.0816] | 13.1271 ^[13.0765] | 17.0499 ^[16.9655] | 21.2152 ^[20.9298] | 9.9175 ^[9.9175] | 21.2103 ^[20.9249] | 0.7619 ^[0.7666] |

Table 7

Component ablation. The numerical values in the top-right corners of “Average (-TSRAG)” and “Average (-KCFT)” indicate the decrease in model performance after ablating those respective components.

| Backbone | Model | GOOGLE BLEU** | BLEU** | | | | | ROUGE** | | | BertScore** |
|--------------|-------------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|---------------|
| | | | BLEU_1** | BLEU_2** | BLEU_3** | BLEU_4** | Mean_BLEU** | ROUGE_1** | ROUGE_2** | ROUGE_J** | |
| Baichuan-7b | Base | 0.6978 | 0.6869 | 0.3373 | 0.2232 | 0.1584 | 0.3515 | 0.3093 | 0.0000 | 0.3093 | 0.5211 |
| | +KCFT | 7.6181 | 9.0244 | 5.0347 | 3.4604 | 2.5064 | 5.0065 | 2.6555 | 0.3307 | 2.6412 | 0.7412 |
| | +TSRAG | 5.8289 | 6.2679 | 4.4854 | 3.7844 | 3.3251 | 4.4657 | 4.7489 | 2.6456 | 4.7489 | 0.5952 |
| | +KCFT+TSRAG(Ours) | 42.8386 | 26.2967 | 22.2469 | 19.6876 | 17.6495 | 21.4702 | 23.9533 | 13.8807 | 23.9533 | 0.7706 |
| Baichuan-13b | Base | 0.8785 | 1.6810 | 1.0381 | 0.7260 | 0.5234 | 0.9921 | 1.0127 | 0.0000 | 1.0127 | 0.5537 |
| | +KCFT | 8.3407 | 8.9717 | 5.0634 | 3.5073 | 2.5425 | 5.0212 | 3.2145 | 0.6137 | 3.2145 | 0.7476 |
| | +TSRAG | 4.9406 | 4.8173 | 2.8906 | 2.2303 | 1.8241 | 2.9406 | 3.7975 | 1.6171 | 3.7975 | 0.5921 |
| | +KCFT+TSRAG(Ours) | 46.6057 | 23.5236 | 20.4608 | 18.5148 | 16.9417 | 19.8602 | 23.0088 | 15.3180 | 23.0088 | 0.7641 |
| ChatGLM3-6B | Base | 0.2083 | 0.0448 | 0.0239 | 0.0160 | 0.0112 | 0.0240 | 0.2743 | 0.0000 | 0.2743 | 0.4916 |
| | +KCFT | 7.4640 | 8.2544 | 4.4424 | 3.0198 | 2.1809 | 4.4744 | 2.1445 | 0.1649 | 2.1345 | 0.7274 |
| | +TSRAG | 8.2231 | 7.2707 | 5.7617 | 5.0692 | 4.5263 | 5.6570 | 6.7411 | 4.3395 | 6.7411 | 0.6127 |
| | +KCFT+TSRAG(Ours) | 19.0147 | 13.6748 | 10.9761 | 9.3027 | 8.0220 | 10.4939 | 12.8622 | 6.3277 | 12.8622 | 0.7359 |
| Llama2-7B | Base | 0.0212 | 0.0174 | 0.0047 | 0.0025 | 0.0015 | 0.0065 | 0.0093 | 0.0000 | 0.0093 | 0.4147 |
| | +KCFT | 6.8692 | 8.5654 | 4.7199 | 3.2239 | 2.3262 | 4.7088 | 2.4301 | 0.1886 | 2.4301 | 0.7315 |
| | +TSRAG | 0.1089 | 0.0083 | 0.0040 | 0.0026 | 0.0019 | 0.0042 | 0.2090 | 0.0262 | 0.2090 | 0.4326 |
| | +KCFT+TSRAG(Ours) | 19.7726 | 11.6220 | 9.1452 | 7.8838 | 7.0323 | 8.9208 | 9.2546 | 5.7006 | 9.2211 | 0.7264 |
| Llama2-13B | Base | 0.0201 | 0.0118 | 0.0042 | 0.0024 | 0.0016 | 0.0050 | 0.0147 | 0.0000 | 0.0147 | 0.4152 |
| | +KCFT | 7.2988 | 8.9108 | 5.0641 | 3.5001 | 2.5494 | 5.0061 | 2.8182 | 0.2759 | 2.8182 | 0.7337 |
| | +TSRAG | 0.1026 | 0.0537 | 0.0437 | 0.0409 | 0.0392 | 0.0444 | 0.3961 | 0.1556 | 0.3961 | 0.4295 |
| | +KCFT+TSRAG(Ours) | 22.8906 | 17.8026 | 13.8248 | 11.6491 | 10.0199 | 13.3241 | 14.5396 | 7.0794 | 14.5396 | 0.7540 |
| Qwen1.5-7B | Base | 0.1440 | 0.0334 | 0.0187 | 0.0128 | 0.0091 | 0.0185 | 0.2814 | 0.0000 | 0.2814 | 0.4962 |
| | +KCFT | 7.4642 | 8.8634 | 5.0621 | 3.4815 | 2.5279 | 4.9837 | 2.9282 | 0.2813 | 2.9282 | 0.7312 |
| | +TSRAG | 2.3810 | 4.5013 | 2.3163 | 1.6635 | 1.2936 | 2.4437 | 1.3295 | 0.3581 | 1.3295 | 0.5767 |
| | +KCFT+TSRAG(Ours) | 50.1202 | 28.7252 | 24.5268 | 21.6901 | 19.5137 | 23.6140 | 27.7619 | 15.9170 | 27.7619 | 0.7799 |
| Qwen1.5-14B | Base | 0.2230 | 0.0436 | 0.0230 | 0.0154 | 0.0108 | 0.0232 | 0.2620 | 0.0000 | 0.2620 | 0.5048 |
| | +KCFT | 7.6993 | 9.0669 | 5.2242 | 3.6188 | 2.6285 | 5.1346 | 3.3072 | 0.3699 | 3.2829 | 0.7362 |
| | +TSRAG | 3.4650 | 4.9997 | 2.7747 | 2.0279 | 1.5670 | 2.8423 | 2.2509 | 0.6332 | 2.2509 | 0.5892 |
| | +KCFT+TSRAG(Ours) | 34.6427 | 22.0391 | 18.0749 | 15.7333 | 13.9787 | 17.4565 | 19.7939 | 11.5174 | 19.7939 | 0.7741 |
| HuaTuo | Base | 0.0951 | 0.7030 | 0.3886 | 0.2607 | 0.1810 | 0.3833 | 0.3083 | 0.0000 | 0.3083 | 0.4948 |
| | +KC | 0.6961 | 2.3882 | 1.0911 | 0.7071 | 0.5013 | 1.1719 | 0.1279 | 0.0000 | 0.1279 | 0.5729 |
| | +TSRAG | 2.1115 | 12.5195 | 10.8965 | 9.7273 | 8.7046 | 10.4620 | 13.2593 | 8.6189 | 13.2593 | 0.6115 |
| | +KC+TSRAG(Ours) | 0.2821 | 32.9406 | 23.5210 | 16.7186 | 11.8590 | 21.2598 | 38.5476 | 3.5994 | 38.5419 | 0.7905 |
| | | | | | | | | | | | |

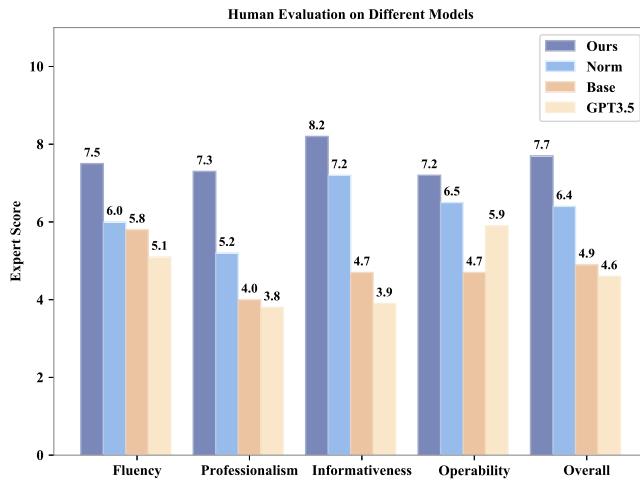


Fig. 6. Human evaluation of outputs from various models. Scores range from 1 to 10.

or disease-related tokens, which require higher precision. Examples can be found in Appendix F.

4.6.2. Human evaluation of models' responses

Same as Section 4.5.3, we conduct a manual evaluation of the domain-specific Q&A generated by the models in terms of fluency, professionalism, information richness, operability, and overall score. This evaluation is intended to demonstrate the effectiveness and feasibility of our knowledge-assimilative LLMs for agricultural decision-making from a professional perspective. For 100 randomly sampled test queries, we employ each optimized LLM for text generation and engage three agricultural experts to evaluate the outputs in a blind trial. The manual scores are shown in Fig. 6, and the Fleiss' Kappa is 0.82. The average scores of LLMs optimized by KALLM across five evaluation criteria are all the highest than other methods, earning the endorsement of agricultural experts. This testament underscores not only the superiority of our approach but also its alignment with the standards of domain expertise.

4.7. Ablation studies

To illustrate the effectiveness of each component within our proposed method and explore their individual contributions, we conduct ablation studies on the two main modules, detailed in Section 4.7.1. Additionally, to thoroughly investigate the model's strengths and areas for development, we perform experimental analyses on the impact of knowledge sources (Section 4.7.4), different fine-tuning strategies (Section 4.7.2), and RAG methods (Section 4.7.3).

4.7.1. Fine-tuning + RAG > fine-tuning only + RAG only

By ablating the KCFT, the model's performance dips significantly across automatic evaluation metrics by 26.1257, 17.0233, 14.2005, 12.0792, 10.4669, 13.4425, 17.1237, 7.6183, 17.1188, and 20.70%, as shown in Table 7. Ablation of the TSRAG results in performance drops of 22.8396, 14.0724, 13.3843, 12.0826, 10.9067, 12.6115, 18.7620, 9.6394, 18.7639, and 4.67%. These reductions confirm the pivotal roles of both KCFT and TSRAG in bolstering the model's domain confidence for decision-making tasks. Notably, the results reveal that, in domain-specific Q&A tasks, mere external knowledge augmentation proves less effective than KCFT alone. This is because general LLMs lack semantic understanding and language generation capabilities within specialized domains. Even with explicit external knowledge, the models are unable to comprehend and may struggle to integrate it

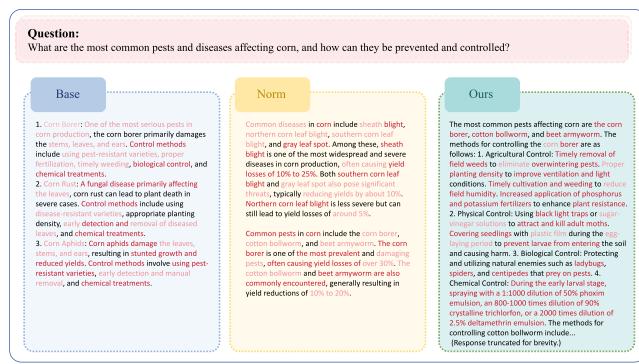


Fig. 7. Comparison of token generation and associated confidence for an agricultural query across different models. “Base” represents the backbone model, “Norm” denotes the model trained using the traditional pipeline, and “Ours” refers to the model trained with our proposed framework. Darker token colours reflects a higher generation confidence.

effectively. In contrast, KCFT assimilates domain knowledge into the model parameters, enabling the model to generate intuitive and albeit approximate responses based on the model's own understanding, even without explicitly external knowledge.

To sum up, we first assimilate knowledge into the model parameters through KCFT and then supplement it with knowledge augmentation, which will help the general LLMs adapt to the specific domain faster. This synergy, where the model's foundational understanding of domain-specific generation is augmented with external knowledge, exemplifies a “1+1>2” effect, offering valuable insights for the future development of domain-specific decision support of LLMs.

4.7.2. Knowledge-coordinated fine-tuning vs. traditional fine-tuning

We compare our knowledge-coordinated fine-tuning with the traditional fine-tuning method to demonstrate its effectiveness in domain-specific knowledge. The comparison results are presented in Table 8.

The results reveal that KCFT outperforms traditional fine-tuning methods across the evaluated metrics, with average improvements of 0.5885, 0.4035, 0.2843, 0.2135, 0.1604, 0.2655, 0.2843, 0.0127, 0.2835, and 0.99%. This preliminary validation suggests adjusting the loss of the corresponding token helps to alter the model's focus on domain-specific terms. Fig. 7 illustrates the confidence levels (probability of token generation during inference) of different models in the domain-specific responses. It is apparent that our model generates high domain confidence for keywords that are closely related to the agricultural domain and provide more informative responses. This further proves the promising application prospects of our proposed framework in agricultural decision-making guidance.

Additionally, we find that fine-tuning HuaTuo with agricultural data, both through traditional and knowledge-coordinated fine-tuning, results in a performance gap compared to other LLMs. Initial analysis suggests this is because HuaTuo has already been trained on medical data, which causes its parameters to shift towards medical knowledge. Compared to other general models, it is more challenging for HuaTuo to adapt to agricultural knowledge.

4.7.3. Comparison of different retrievers

To investigate the impact of different retrievers on model performance, we conduct an ablation study on the retriever used in TSRAG. Specifically, we replace the topic-matched retriever with DPR (trained on the dataset), a general BERT-based retriever, and BM25. The comparative outcomes are delineated in Table 9. The results illustrate that character-matching sparse retrieval methods outperform dense retrieval methods based on semantic encoding similarity in specific domains. This is because the semantics of Q&A pairs are pretty similar, making it challenging to find precise knowledge entries (e.g., rice blast and rice

Table 8

Knowledge-coordinated fine-tuning vs. traditional fine-tuning. The number in the top-right corner of “Average (KCFT)” represents the improvement achieved by the model when utilizing KC fine-tuning over traditional fine-tuning.

| Backbone | Data | GOOGLE BLEU** | BLEU** | | | | | ROUGE** | | | BertScore** |
|--------------|------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|--------------------------------|
| | | | BLEU_1** | BLEU_2** | BLEU_3** | BLEU_4** | Mean_BLEU** | ROUGE_1** | ROUGE_2** | ROUGE_L** | |
| Baichuan-7b | Norm | 7.4132 | 9.3067 | 5.2654 | 3.6148 | 2.6195 | 5.2016 | 2.8009 | 0.5649 | 2.8009 | 0.7409 |
| | KCFT | 7.6181 | 9.0244 | 5.0347 | 3.4604 | 2.5064 | 5.0065 | 2.6555 | 0.3307 | 2.6412 | 0.7412 |
| Baichuan-13b | Norm | 4.9808 | 7.4960 | 4.1315 | 2.8545 | 2.0789 | 4.1402 | 1.9056 | 0.0556 | 1.9056 | 0.7243 |
| | KCFT | 8.3407 | 8.9717 | 5.0634 | 3.5073 | 2.5425 | 5.0212 | 3.2145 | 0.6137 | 3.2145 | 0.7476 |
| ChatGLM3-6B | Norm | 7.0400 | 8.9245 | 4.9094 | 3.3419 | 2.4130 | 4.8972 | 2.6041 | 0.2041 | 2.6041 | 0.7305 |
| | KCFT | 7.4640 | 8.2544 | 4.4424 | 3.0198 | 2.1809 | 4.4744 | 2.1445 | 0.1649 | 2.1345 | 0.7274 |
| Llama2-7B | Norm | 6.8395 | 8.3513 | 4.4884 | 3.0034 | 2.1515 | 4.4986 | 2.0858 | 0.1895 | 2.0706 | 0.7266 |
| | KCFT | 6.8692 | 8.5654 | 4.7199 | 3.2239 | 2.3262 | 4.7088 | 2.4301 | 0.1886 | 2.4301 | 0.7315 |
| Llama2-13B | Norm | 7.3336 | 8.6182 | 4.6256 | 3.1320 | 2.2562 | 4.6580 | 2.3725 | 0.3577 | 2.3590 | 0.7274 |
| | KCFT | 7.2988 | 8.9108 | 5.0641 | 3.5001 | 2.5494 | 5.0061 | 2.8182 | 0.2759 | 2.8182 | 0.7337 |
| Qwen1.5-7B | Norm | 7.5956 | 8.8390 | 4.8991 | 3.3732 | 2.4474 | 4.8897 | 2.4759 | 0.3370 | 2.4561 | 0.7313 |
| | KCFT | 7.4642 | 8.8634 | 5.0621 | 3.4815 | 2.5279 | 4.9837 | 2.9282 | 0.2813 | 2.9282 | 0.7312 |
| Qwen1.5-14B | Norm | 7.4353 | 8.8350 | 4.8473 | 3.3147 | 2.3905 | 4.8469 | 2.8454 | 0.4149 | 2.8454 | 0.7368 |
| | KCFT | 7.6993 | 9.0669 | 5.2242 | 3.6188 | 2.6285 | 5.1346 | 3.3072 | 0.3699 | 3.2829 | 0.7362 |
| HuaTuo | Norm | 0.1048 | 0.4461 | 0.2605 | 0.1765 | 0.1226 | 0.2514 | 0.2612 | 0.0000 | 0.2612 | 0.5250 |
| | KCFT | 0.6961 | 2.3882 | 1.0911 | 0.7071 | 0.5013 | 1.1719 | 0.1279 | 0.0000 | 0.1217 | 0.5729 |
| Average | Norm | 6.0929 | 7.6021 | 4.1784 | 2.8514 | 2.0600 | 4.1730 | 2.1689 | 0.2655 | 2.1629 | 0.7053 |
| | KCFT | 6.6813^{±0.5885} | 8.0057^{±0.4035} | 4.4627^{±0.2843} | 3.0649^{±0.2135} | 2.2204^{±0.1604} | 4.4384^{±0.2655} | 2.4533^{±0.2843} | 0.2781^{±0.0127} | 2.4464^{±0.2835} | 0.7152^{±0.99%} |

Table 9

Comparison of different retrievers. The numeral in the top-right corner of “Average (TSRAG)” indicates the enhanced performance of the model when utilizing TSRAG compared to BM25.

| Backbone | Retriever | GOOGLE BLEU** | BLEU** | | | | | ROUGE** | | | BertScore** |
|--------------|-----------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|--------------------------------|
| | | | BLEU_1** | BLEU_2** | BLEU_3** | BLEU_4** | Mean_BLEU** | ROUGE_1** | ROUGE_2** | ROUGE_L** | |
| Baichuan-7b | DPR | 7.0585 | 8.0242 | 4.4293 | 3.0054 | 2.1666 | 4.4064 | 2.1580 | 0.2033 | 2.1294 | 0.7015 |
| | Bert | 5.0562 | 7.1662 | 4.1111 | 2.8734 | 2.0920 | 4.0607 | 2.0140 | 0.0661 | 2.0140 | 0.6933 |
| | BM25 | 45.9595 | 19.8551 | 17.5801 | 16.2682 | 15.1411 | 17.2111 | 17.6403 | 13.2503 | 17.6147 | 0.7522 |
| Baichuan-13b | TSRAG | 42.8386 | 26.2967 | 22.2469 | 19.6876 | 17.6495 | 21.4702 | 23.9533 | 13.8807 | 23.9533 | 0.7706 |
| | DPR | 1.5236 | 2.2124 | 1.2557 | 0.8648 | 0.6184 | 1.2378 | 1.0278 | 0.0000 | 1.0278 | 0.5681 |
| | Bert | 2.1475 | 3.3833 | 2.2519 | 1.6622 | 1.2421 | 2.1349 | 1.3951 | 0.2500 | 1.3951 | 0.5902 |
| | BM25 | 34.7047 | 12.0909 | 10.2067 | 8.9963 | 8.1478 | 9.8604 | 11.3037 | 6.4984 | 11.3037 | 0.7613 |
| ChatGLM3-6B | TSRAG | 46.6057 | 23.5236 | 20.4608 | 18.5148 | 16.9417 | 19.8602 | 23.0088 | 15.3180 | 23.0088 | 0.7641 |
| | DPR | 6.7321 | 8.1606 | 4.5012 | 3.0932 | 2.2385 | 4.4984 | 2.1695 | 0.1801 | 2.1593 | 0.6999 |
| | Bert | 3.8788 | 5.9831 | 3.4440 | 2.3986 | 1.7511 | 3.3942 | 1.4963 | 0.0294 | 1.4816 | 0.6455 |
| | BM25 | 13.4720 | 10.0660 | 7.4536 | 6.0671 | 5.0602 | 7.1617 | 8.2549 | 3.7044 | 8.2549 | 0.7273 |
| Llama2-7B | TSRAG | 19.0147 | 13.6748 | 10.9761 | 9.3027 | 8.0220 | 10.4939 | 12.8622 | 6.3277 | 12.8622 | 0.7359 |
| | DPR | 6.3113 | 7.7408 | 4.1779 | 2.8437 | 2.0529 | 4.2038 | 2.0861 | 0.1102 | 2.0726 | 0.7001 |
| | Bert | 4.8936 | 7.2036 | 3.9774 | 2.7197 | 1.9674 | 3.9670 | 1.6500 | 0.0955 | 1.6500 | 0.6988 |
| | BM25 | 16.6862 | 14.4359 | 11.1381 | 9.1851 | 7.7974 | 10.6391 | 12.5245 | 5.4874 | 12.5245 | 0.7432 |
| Llama2-13B | TSRAG | 19.7726 | 11.6220 | 9.1452 | 7.8838 | 7.0323 | 8.9208 | 9.2546 | 5.7006 | 9.2211 | 0.7264 |
| | DPR | 6.9563 | 8.1111 | 4.4013 | 2.9588 | 2.1184 | 4.3974 | 2.3822 | 0.2369 | 2.3675 | 0.7026 |
| | Bert | 5.6650 | 7.6821 | 4.3412 | 2.9738 | 2.1523 | 4.2873 | 2.4249 | 0.1268 | 2.4201 | 0.7044 |
| | BM25 | 23.1690 | 13.6940 | 11.0128 | 9.3876 | 8.2167 | 10.5778 | 12.6972 | 6.5239 | 12.6972 | 0.7400 |
| Qwen1.5-7B | TSRAG | 22.8906 | 17.8026 | 13.8248 | 11.6491 | 10.0199 | 13.3241 | 14.5396 | 7.0794 | 14.5396 | 0.7540 |
| | DPR | 6.9883 | 8.8782 | 4.7718 | 3.2299 | 2.3241 | 4.8010 | 2.7802 | 0.5478 | 2.7713 | 0.7110 |
| | Bert | 6.8567 | 7.7186 | 4.2355 | 2.8669 | 2.0609 | 4.2205 | 2.1314 | 0.2319 | 2.1261 | 0.7040 |
| | BM25 | 43.1099 | 22.9652 | 20.1930 | 18.3995 | 16.9792 | 19.6342 | 22.6264 | 15.7566 | 22.6264 | 0.7630 |
| Qwen1.5-14B | TSRAG | 50.1202 | 28.7252 | 24.5268 | 21.6901 | 19.5137 | 23.6140 | 27.7619 | 15.9170 | 27.7619 | 0.7799 |
| | DPR | 7.9021 | 8.0816 | 4.4448 | 3.0423 | 2.2019 | 4.4427 | 2.4151 | 0.3434 | 2.3999 | 0.7113 |
| | Bert | 6.4287 | 8.4999 | 4.5555 | 3.0905 | 2.2241 | 4.5925 | 1.9265 | 0.1004 | 1.9265 | 0.7080 |
| | BM25 | 25.6375 | 16.3840 | 12.7805 | 10.9027 | 9.4431 | 12.3776 | 12.8866 | 7.2493 | 12.8866 | 0.7461 |
| Average | TSRAG | 34.6427 | 22.0391 | 18.0749 | 15.7333 | 13.9787 | 17.4565 | 19.7939 | 11.5174 | 19.7939 | 0.7741 |
| | DPR | 6.2103 | 7.3156 | 3.9974 | 2.7197 | 1.9601 | 3.9982 | 2.1455 | 0.2317 | 2.1325 | 0.6849 |
| | Bert | 4.9895 | 6.8053 | 3.8452 | 2.6550 | 1.9271 | 3.8082 | 1.8626 | 0.1286 | 1.8590 | 0.6778 |
| | BM25 | 28.9627 | 15.6416 | 12.9093 | 11.3152 | 10.1122 | 12.4946 | 13.9905 | 8.3529 | 13.9869 | 0.7476 |
| | TSRAG | 33.6979^{±4.7352} | 20.5263^{±4.8847} | 17.0363^{±4.1272} | 14.9231^{±3.6078} | 13.3083^{±3.1690} | 16.4485^{±3.9540} | 18.7392^{±4.7486} | 10.8201^{±2.4672} | 18.7344^{±4.7475} | 0.7578^{±1.03%} |

sheath blight differ only in the disease name, with similar semantics). To substantiate this point, we sample different categories of knowledge from the knowledge base and visualize them in the semantic vector space. The Fig. 8-a shows that knowledge entries in the domain-specific knowledge base are semantically similar.

Additionally, DPR trained on domain data performs better than BERT. It further demonstrates that dense retrieval using text embeddings is limited by the accuracy of semantic encoding. Conversely, sparse retrievers may not deeply understand the semantics of tokens, but they can achieve higher accuracy in specific domain scenarios with highly similar text.

We quantitatively evaluate the accuracy of the evidence retrieved using different retrievers, as shown in Fig. 8-b. The evidence accuracy aligns with our hypothesis, highlighting the nuanced strengths of our method in handling domain-specific knowledge retrieval.

4.7.4. The impact of different retrieval sources on retrieval results

Taking into account the domain relevance and factual consistency requirements of the external knowledge base, we leverage books written by agricultural experts as high-quality domain-specific sources. To validate the superiority of our knowledge base in furnishing semantically intensive information, we conduct comparative experiments with commonly web-based repositories, including Google, Wikipedia, and ChatGPT, as alternative knowledge sources. The experiments include manual assessments of retrieval results and automated quantitative evaluations of generated responses based on the retrieved information. The experimental results are shown in Fig. 9. Examples of the retrieval results comparison are provided in Appendix E.

The effectiveness of external knowledge augmentation is significantly improved when the evidence retrieved is highly relevant to the query. Meanwhile, when the provided evidence is more reliable, the generated responses yield greater domain confidence. As can be seen from Fig. 9, our curated knowledge base demonstrates superior specialization and informativeness, whether assessed through automated metrics or expert scrutiny. Therefore, we believe that professional books as a reliable source can provide enhanced knowledge guidance for models to generate more domain-specific outputs.

5. Limitation and future work

Under the guidance of agricultural experts, we craft annotation guidelines and establish a large-scale agricultural dataset for four pivotal tasks: crop variety recommendation, cultivation management, crop pest and disease diagnosis and prevention, alongside compiling relevant Q&A pairs and a knowledge base. However, it is noteworthy that our model, based on the format of Q&A pairs, only handles single-turn agricultural queries while cannot implement multi-turn dialogues. To study dialogue historical memory and context-fusion understanding, we will focus on building multi-turn agricultural dialogue datasets for model training in the future.

We also find that the imbalance of data affects the knowledge injection of the model: the larger the dataset, the more beneficial the knowledge infusion. To ensure substantial model improvement in tasks, it is crucial to acquire more relevant knowledge and Q&A pairs. This will motivate us to expand the related datasets further.

Furthermore, our model operates in a single-modal manner, relying solely on text inputs for decision guidance. This poses challenges, especially in tasks such as pest diagnosis, where inaccurate descriptions could affect accuracy. Incorporating multimodality (i.e., text and image) is a future frontier for improving decision precision and matching farmers' actual needs.

Additionally, all the knowledge used in this study is textual, which may introduce issues of high redundancy and low efficiency during evidence retrieval. As a clear semantic expression and knowledge structure, knowledge graphs can produce interpretable reasoning paths to enhance the model's transparency during knowledge constraint processes. Consequently, we are exploring the construction of agricultural knowledge graphs to assist the semantic reasoning of LLM.

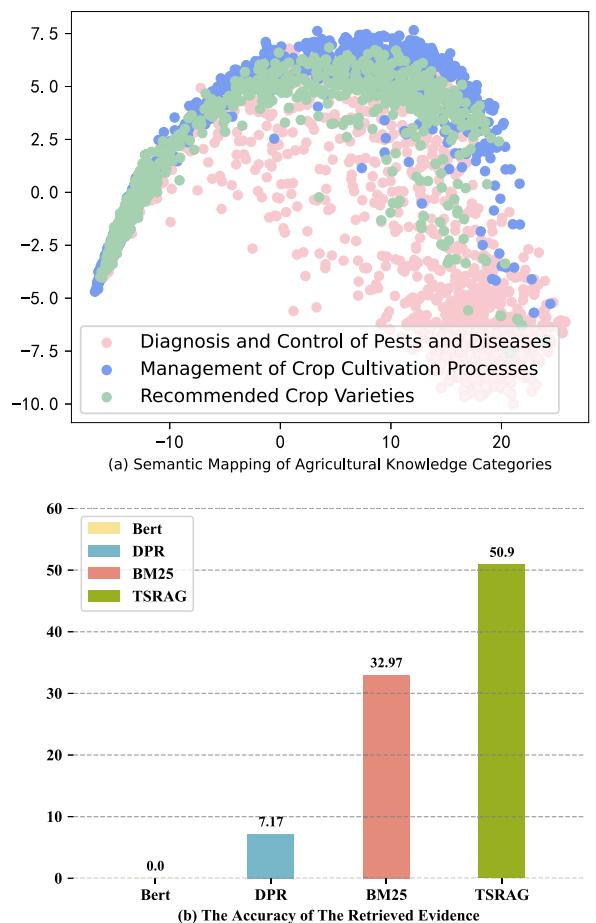


Fig. 8. Semantic space of knowledge entries and accuracy of different retrievers. Various types of knowledge from the knowledge base are encoded semantically via BERT and mapped to a two-dimensional space.

6. Conclusion

In this study, we develop a knowledge-guided LMM for agricultural decision-making tasks. To address the shortcomings of high-quality data and the high-confidence decision-making model in agriculture, our work focuses on building a pipeline of data annotation, implicit injection of domain knowledge, and explicit augmentation of external knowledge. An expert-annotated corpus consisting of 220,000 Q&A pairs and a corresponding knowledge base of more than 15 million tokens are first presented in the field of agriculture. Meanwhile, we provide valuable insights for the dataset construction of similar domain-specific LLMs in terms of data sources, data quality, data size, data evaluation, etc. On the basis of these high-quality data, we experimentally verify that the proposed KALLM method incorporating knowledge-coordinated fine-tuning and self-reflection retrieval-augment generation can improve the agricultural decision-making ability of eight general LLMs in all 10 metrics. A large number of ablation experiments and practical examples also significantly prove that our token-level supervised fine-tuning method and sentence-level retrieval-augment generation method are necessary in the process of transferring general LLMs to domain-specific LLMs. Furthermore, from our experimental analysis, we have drawn the following intriguing conclusions:

- Constructing data or knowledge bases from professional books is more effective than sourcing them from the internet, as the latter often contains significant redundancy.

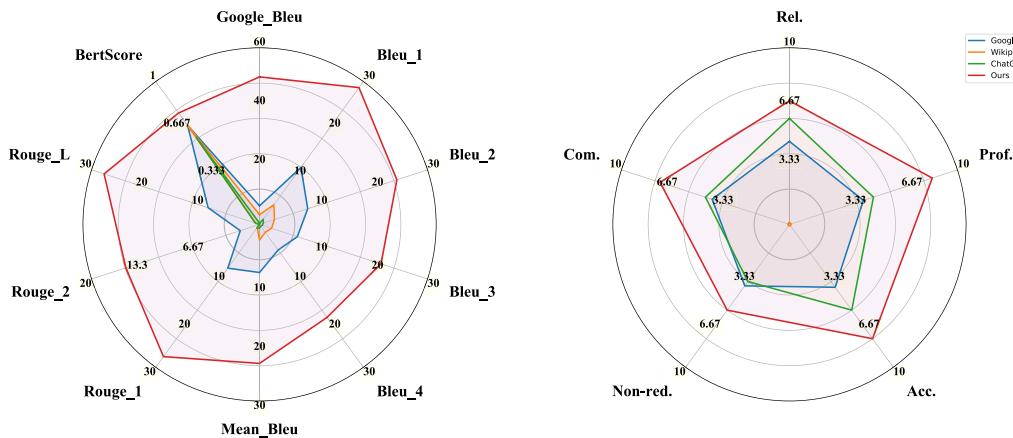


Fig. 9. Comparison of different retrieval sources. Quantitative evaluation of generated responses based on retrieval results from four knowledge bases (Left), along with ratings from agricultural experts (Right).

- Larger datasets lead to more knowledge injected into the model and better performance. However, we have not explored whether this rule always holds (i.e., whether the scaling law fails when the dataset exceeds a certain magnitude).
- Data imbalance can adversely affect task performance, with larger datasets generally leading to better outcomes.
- Adjusting model prediction weights to amplify the contribution of certain tokens to the loss indeed enhances the model's focus on these tokens.
- In knowledge bases with only a few token differences and highly similar semantics, sparse retrieval through character matching proves more effective than dense retrieval based on semantic similarity.

We believe that the agricultural knowledge-guided architecture proposed in this paper can provide an essential reference for the construction of domain-specific LLMs.

CRediT authorship contribution statement

Jingchi Jiang: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Lian Yan:** Writing – review & editing, Validation, Supervision, Software, Resources. **Haifeng Liu:** Validation, Software, Investigation, Data curation. **Zhenbo Xia:** Validation, Software, Investigation. **Haotian Wang:** Visualization, Data curation. **Yang Yang:** Validation, Data curation. **Yi Guan:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study was supported in part by grants from the National Science and Technology Major Project [ZDYZF20220008-02], the National Key R&D Program of China [2021ZD0110900], and the National Natural Science Foundation of China [62006063]. We thank all the anonymous reviewers who generously contributed their time and efforts. Their professional recommendations have greatly enhanced the quality of the manuscript.

Appendix A. Evaluation metrics

• BLEU

$$BLEU_N = BP \cdot \exp\left(\sum_{n=1}^N w_n \log p_n\right) \quad (\text{A.1})$$

If Y denotes the reference answer, with gram_n^Y as its decomposition into N-grams, A is the model's predicted output, decomposed into N-grams as gram_n^A , $\text{Count}(Y)$ and $\text{Count}(A)$ represent the lengths of Y and A , respectively, and $\text{Count}(a, b)$ the number of times a appears in b , then

$$BP = \begin{cases} 1 & \text{if } \text{Count}(A) > \text{Count}(Y) \\ \exp\left(1 - \frac{\text{Count}(Y)}{\text{Count}(A)}\right) & \text{else } \text{Count}(A) \leq \text{Count}(Y), \end{cases} \quad w_n = \frac{1}{N},$$

$$p_n = \frac{\sum_{i \in \text{gram}_n^A} \min(\text{Count}(i, \text{gram}_n^A), \text{Count}(i, \text{gram}_n^Y))}{\sum_{i \in \text{gram}_n^A} \text{Count}(i, \text{gram}_n^A)}.$$

• Google_Bleu

Differing from $BLEU_N$, $Google_Bleu$ incorporates a smoothing factor θ into its p_n calculation, defined as: $p_n = \frac{\sum_{i \in \text{gram}_n^A} \min(\text{Count}(i, \text{gram}_n^A), \text{Count}(i, \text{gram}_n^Y)) + \theta}{\sum_{i \in \text{gram}_n^A} \text{Count}(i, \text{gram}_n^A) + \theta}$, where θ is commonly set to 0.1.

• ROUGE

$$Rouge_N = \frac{\sum_{\substack{\text{gram}_n^A \in A \\ \text{gram}_n^Y \in Y}} \text{Count}_{match}(\text{gram}_n^A)}{\sum_{\text{gram}_n^Y \in Y} \text{Count}(\text{gram}_n^Y)} \quad (\text{A.2})$$

where Y denotes the reference answer, gram_n^Y is the result of splitting Y into N-grams, A is the model's predicted output, and gram_n^A is the result of splitting A into N-grams. $\text{Count}(\text{gram}_n^Y)$ denotes the count of gram_n^Y , and $\text{Count}_{match}(\text{gram}_n^A)$ accounts for the tally of matching N-grams found between gram_n^Y and gram_n^A .

$$Rouge_L = \frac{(1 + \alpha^2)R_{lcs}P_{lcs}}{R_{lcs} + \alpha^2 P_{lcs}} \quad (\text{A.3})$$

where $\alpha = 1$, $R_{lcs} = \frac{LCS(Y, A)}{\text{Count}(Y)}$, $P_{lcs} = \frac{LCS(Y, A)}{\text{Count}(A)}$, and $LCS(Y, A)$ represents the longest common subsequence between Y and A .

• BertScore

$$BertScore = \frac{\text{Bert}(Y) \cdot \text{Bert}(A)}{\|\text{Bert}(Y)\| \cdot \|\text{Bert}(A)\|} \quad (\text{A.4})$$

Appendix B. Annotated data samples

We present four data samples in Figs. B.1, B.2, B.3, B.4, and B.5 respectively.

Data Samples1: QA & Knowledge

```
{
id:"XXXXXX",
source:"Pests and Diseases of Crops in China",
```

Knowledge: "Epidemic Patterns of Wheat Dwarf Disease. The gray planthopper is the primary vector for wheat dwarf disease. Planting systems, cultivation management practices, and weather conditions that favor the reproduction of the vector insect and the preservation of the virus contribute to the occurrence of wheat dwarf disease. The disease often appears at the edges of fields, near ditches, and on late autumn crops where weeds are abundant, attracting gray planthoppers. Some weeds are also hosts of the virus, leading to more severe disease in wheat fields near planthopper habitats. Intercropped fields have higher disease incidence, while fields with fine plowing have lower incidence. After harvesting autumn crops, fields with many weeds or where wheat is directly intercropped with autumn crops have higher planthopper populations, leading to more severe disease as the insects feed and transmit the virus to emerging wheat seedlings. Early-sown fields show higher disease incidence, while fields sown at the optimal time show lower incidence. Early-sown wheat emerges early, coinciding with the peak activity period of gray planthoppers before winter, providing more opportunities for infection. The higher temperatures favor virus multiplication and accumulation, leading to severe disease both before and after winter. In years with heavy summer and autumn rainfall, the humid climate promotes weed growth and planthopper reproduction. Warm winters and cold springs favor planthopper overwintering but hinder wheat seedling growth, resulting in more severe disease in years with such weather conditions.",

Forward-Query: "Is the likelihood of wheat dwarf disease higher in years with heavy summer and autumn rainfall, and warm winters followed by cold springs?",

Forward-Answer: "The probability of wheat dwarf disease is relatively high. The gray planthopper is the primary vector for wheat dwarf disease. Planting systems, cultivation management practices, and weather conditions that favor the reproduction of the vector insect and the preservation of the virus contribute to the occurrence of wheat dwarf disease. In years with heavy summer and autumn rainfall, the humid climate promotes weed growth and planthopper reproduction. Warm winters and cold springs favor planthopper overwintering but hinder wheat seedling growth. Therefore, the incidence of wheat dwarf disease is higher in years with such weather conditions.",

Reverse-query: "Under what environmental conditions is there a higher likelihood of wheat rosette dwarf disease?",

Reverse-answer: "The small brown planthopper (*Laodelphax striatellus*) is the main vector for transmitting the virus causing wheat rosette dwarf disease. In years with rainy summers and autumns, the climate is humid, and weeds proliferate, which is beneficial for the reproduction and summer survival of the small brown planthopper. Warm winters followed by cold springs are conducive to the overwintering of the small brown planthopper but unfavorable for the growth and development of wheat seedlings. Therefore, the disease is more severe in years with rainy summers and autumns, and warm winters followed by cold springs."

```
}
```

Fig. B.1. Data Samples1.

Data Samples2: QA & Knowledge

```
{
id:"XXXXXX",
```

source: "National Crop Promotion Catalogue",

Knowledge: "Rice Variety '10 You 18'. Approval Opinion: This variety meets the national standards for rice variety approval and has been approved. It has a moderate maturity period, high yield, excellent grain quality, and resistance to rice blast. It is suitable for planting in the single-crop spring rice regions of Beijing, Tianjin, and the eastern and central-northern parts of Hebei Province. Characteristics: This variety is a japonica type three-line hybrid rice... Yield Performance: In the 2007 regional trials for the Beijing-Tianjin-Tangshan japonica rice group, it had an average yield of 679.2 kg per mu... Cultivation Techniques: 1. Seedling Raising: Sow at the appropriate time. Before sowing, dry the seeds in the sun and disinfect them to prevent dry seedling blight and bakanae disease. Use about half the seed amount of conventional varieties to cultivate robust seedlings with tillers. 2. Transplanting: Transplant seedlings at around 35 days old, with a spacing of 30 cm × 13.3 cm, planting two seedlings per hill...",

Query: "What are the cultivation techniques for the rice variety '10 You 18', suitable for the single-crop spring rice regions of Beijing, Tianjin, and eastern and central-northern Hebei Province?",

Answer: "1. Seedling Raising: Sow at the appropriate time. Before sowing, dry the seeds in the sun and disinfect them to prevent dry seedling blight and bakanae disease. Use about half the seed amount of conventional varieties to cultivate robust seedlings with tillers. 2. Transplanting: Transplant seedlings at around 35 days old, with a spacing of 30 cm × 13.3 cm, planting two seedlings per hill. 3. Fertilizer and Water Management: Use a combination of nitrogen, phosphorus, potassium, and zinc fertilizers. Alternate between wet and dry conditions to ensure about 180,000 effective panicles per mu. 4. Pest and Disease Control: Pay attention to the prevention of false smut. Manage other pests, diseases, and weeds as with conventional rice varieties."

```
}
```

Fig. B.2. Data Samples2.

Data Samples3: QA & Knowledge

```
{
id:"XXXXXX",
```

source: "National Germplasm Website",

Knowledge: "Soybean Variety 'Andou 5156'. Approval Opinion: This variety meets the soybean variety approval standards and has been approved. It is suitable for planting in summer soybean regions across Henan Province. Characteristics: It is a determinate, mid-maturing variety with a growth period of 107.0–112.2 days. The plant has a compact structure, with a height ranging from 62.6 to 78.9 cm and 1.9 to 3.1 effective branches per plant... Yield Performance: In the 2013 Henan Province soybean variety regional trials, it was tested at 12 sites, showing yield increases at 11 sites. The average yield was 205.2 kg per mu... Cultivation Techniques: 1. Sowing Time and Seed Rate: The optimal sowing period is in early to mid-June, with a seed rate of about 5 kg per mu. The row spacing should be 0.4 m, and the plant spacing should be 0.13 m, resulting in a planting density of approximately 12,500 plants per mu...",

Query: "Can you recommend a soybean variety suitable for planting in summer soybean regions across Henan Province?",

Answer: "'Andou 5156'. This variety meets the soybean variety approval standards and has been approved. It is suitable for planting in summer soybean regions across Henan Province."

```
}
```

Fig. B.3. Data Samples3.

| Data Samples4: QA & Knowledge |
|---|
| <pre>{ id:"XXXXXX", source:"Pests and Diseases of Crops in China", Knowledge:"Soybean Moth Control Techniques. (I) Agricultural Control. Selecting Resistant Varieties: The susceptibility of soybean varieties to pests varies. Soybean moth larvae generally prefer early-maturing, soft-stemmed varieties with high protein and fat content. Therefore, selecting late-maturing varieties with hard stems, thick skins, and strong flood resistance can reduce soybean moth damage. Timely Autumn Plowing to Reduce Overwintering Population: After the soybean harvest in the autumn and winter, timely deep plowing can turn the mature larvae overwintering in the soil to the surface or destroy their overwintering sites. The mechanical damage and cold winter weather can kill a significant portion of the pests, reducing the number of adults in the following year. Crop Rotation: Rotating between wet and dry crops and avoiding continuous planting of leguminous plants can mitigate damage. (II) Manual Control. When the larvae reach the fourth instar or older, manual control measures such as hand-picking or cutting with scissors can be employed... (III) Physical Control. Utilizing the moth's strong phototaxis, black light traps can be set up to attract and kill adults, thereby reducing the number of eggs laid in soybean fields and the subsequent feeding damage by larvae. (IV) Biological Control. Using Bacillus thuringiensis or Bacillus entomocidus (with a spore count of 8-10 billion per gram), dilute to 500-700 times the concentration, and apply 50 kg of the solution per 667 m² to control young soybean moth larvae. (V) Chemical Control. The optimal time for chemical control is during the peak period of the first to third instar larvae. Spray when the number of larvae reaches 10 per 100 plants, preferably in the afternoon. The timing of spraying significantly affects the control efficacy, and the uniformity of application is also crucial, particularly on the underside of leaves. Recommended insecticides include 40% Phoxim EC at 1500 times dilution, 2.5% Deltamethrin EC at 3000-4000 times dilution, 4.5% Beta-cypermethrin EC at 1500 times dilution, 20% Fenvalerate EC at 1000-2000 times dilution, 45% Malathion EC at 1000 times dilution, or 25% Diflubenzuron SC at 1000 times dilution.". Query:"How to control soybean moths?", Answer:" (I) Agricultural Control. 1. Selecting Resistant Varieties... Choose late-maturing varieties with hard stems, thick skins, and strong flood resistance to reduce soybean moth damage. 2. Timely Autumn Plowing to Reduce Overwintering Population. After the soybean harvest in the autumn and winter, deep plowing can bring the mature larvae overwintering in the soil to the surface or destroy their overwintering sites... 3. Crop Rotation. Rotate between wet and dry crops and avoid continuous planting of leguminous plants to mitigate damage. (II) Manual Control. When the larvae reach the fourth instar or older, manual control measures such as hand-picking or cutting with scissors can be employed... (III) Physical Control. Utilize the moth's strong phototaxis by setting up black light traps to attract and kill adults, reducing the number of eggs laid and subsequent larval damage. (IV) Biological Control. Use Bacillus thuringiensis or Bacillus entomocidus (with a spore count of 8-10 billion per gram) diluted 500-700 times. Apply 50 kg of the solution per 667 m² to control young larvae. (V) Chemical Control. The optimal time for chemical control is during the peak period of the first to third instar larvae. Spray when the number of larvae reaches 10 per 100 plants, preferably in the afternoon... Ensure uniform application, particularly on the underside of leaves. Recommended insecticides include 40% Phoxim EC at 1500 times dilution, 2.5% Deltamethrin EC at 3000-4000 times dilution, 4.5% Beta-cypermethrin EC at 1500 times dilution..."}</pre> |

Fig. B.4. Data Samples4.

| Annotated Content: QA & Knowledge |
|---|
| <pre>{ id:"XX", #Number id source:"XXXX", #Where the information comes, web pages or books Knowledge:"XXXX", #Knowledge items in a knowledge base, derived from fragments in books. Query:"XXXX", #Question constructed from corresponding pieces of knowledge. Answer:"XXXX" #The answer to the question. }</pre> |

Fig. B.5. Specific content of annotated data.

Appendix C. Comparison examples of model performance after fine-tuning with different datasets

The comparison examples of model performance after fine-tuning with the manually annotated dataset and the automatically constructed dataset can be seen in Fig. C.1.

The agricultural experts score the model responses fine-tuned with two different datasets based on their relevance to the query, professionalism, and alignment with the reference answers. Scores ranged from 1 to 10, and the results are shown in Fig. C.2.

Appendix D. Responses generated by different models

Fig. D.1 showcases responses generated by different models.

Appendix E. Retrieval evidence obtained from different knowledge sources

The evidence obtained from different knowledge sources is shown in Fig. E.1.

Appendix F. The models' performance on the four tasks

Table F.1 shows the models' performance on the four tasks. The query and retrieved knowledge of Crop Variety Recommendation example can be seen in Fig. F.1.

Data availability

Our code and data are available at <https://github.com/YanPioneer/KALLM>.

Table F.1
The models' performance on the four tasks.

| Model | Task | GOOGLE | BLEU** | | | | | ROUGE** | | | BertScore** |
|-------------------------|----------------------------------|---------|----------|----------|----------|----------|-------------|-----------|-----------|-----------|-------------|
| | | BLEU** | BLEU_1** | BLEU_2** | BLEU_3** | BLEU_4** | Mean_BLEU** | ROUGE_1** | ROUGE_2** | ROUGE_1** | |
| GPT3.5 (TSRAG) | Crop Variety Recommendation | 0.0013 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.4457 |
| | Cultivation Management | 1.8006 | 10.7886 | 5.8401 | 4.1059 | 3.0535 | 5.9470 | 5.9836 | 1.9895 | 5.9836 | 0.6834 |
| | Crop Pest and Disease Diagnosis | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.5444 |
| | Crop Pest and Disease Prevention | 0.2809 | 1.9973 | 1.4860 | 1.1103 | 0.8340 | 1.3569 | 0.0000 | 0.0000 | 0.0000 | 0.6297 |
| GPT4 (TSRAG) | Crop Variety Recommendation | 0.0083 | 0.1711 | 0.1210 | 0.0856 | 0.0605 | 0.1096 | 0.3101 | 0.0000 | 0.3101 | 0.4676 |
| | Cultivation Management | 3.6966 | 10.6678 | 6.0049 | 4.4332 | 3.4345 | 6.1351 | 8.8097 | 1.5245 | 8.8097 | 0.8002 |
| | Crop Pest and Disease Diagnosis | 0.0000 | 0.0282 | 0.0199 | 0.0141 | 0.0100 | 0.0181 | 0.0000 | 0.0000 | 0.0000 | 0.5070 |
| | Crop Pest and Disease Prevention | 1.3285 | 2.1233 | 0.8562 | 0.5144 | 0.3400 | 0.9585 | 0.6721 | 0.1669 | 0.6721 | 0.7103 |
| Claude3.5 (TSRAG) | Crop Variety Recommendation | 0.0080 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0182 | 0.0000 | 0.0182 | 0.4444 |
| | Cultivation Management | 2.3947 | 2.9941 | 1.5868 | 1.0354 | 0.7016 | 1.5795 | 5.0814 | 0.7562 | 5.0814 | 0.7633 |
| | Crop Pest and Disease Diagnosis | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.4969 |
| | Crop Pest and Disease Prevention | 1.6549 | 2.1491 | 1.0815 | 0.7248 | 0.5088 | 1.1160 | 0.6580 | 0.1166 | 0.6580 | 0.6867 |
| Baichuan-7b (KALLM) | Crop Variety Recommendation | 6.7925 | 7.4419 | 5.3984 | 3.8575 | 2.7432 | 4.8603 | 7.4419 | 0.4651 | 7.4419 | 0.6849 |
| | Cultivation Management | 50.1832 | 50.0302 | 43.3867 | 39.4890 | 36.2607 | 42.2917 | 44.8586 | 30.5410 | 44.8586 | 0.8773 |
| | Crop Pest and Disease Diagnosis | 0.3421 | 9.2308 | 6.5271 | 4.6154 | 3.2636 | 5.9092 | 9.2308 | 0.0000 | 9.2308 | 0.6682 |
| | Crop Pest and Disease Prevention | 11.5120 | 20.4244 | 18.1769 | 16.8845 | 15.5468 | 17.7582 | 16.3851 | 13.4524 | 16.3851 | 0.7223 |
| Baichuan-13b (KALLM) | Crop Variety Recommendation | 3.2258 | 3.7209 | 2.6311 | 1.8605 | 1.3155 | 2.3820 | 3.7209 | 0.0000 | 3.7209 | 0.6674 |
| | Cultivation Management | 53.4928 | 48.3036 | 42.6823 | 39.2131 | 36.3228 | 41.6305 | 47.3052 | 34.4940 | 47.3052 | 0.8827 |
| | Crop Pest and Disease Diagnosis | 0.1010 | 1.5385 | 1.0879 | 0.7692 | 0.5439 | 0.9849 | 1.5385 | 0.0000 | 1.5385 | 0.6477 |
| | Crop Pest and Disease Prevention | 17.1107 | 19.5412 | 18.6565 | 17.9506 | 17.1686 | 18.3292 | 15.6381 | 11.3459 | 15.6381 | 0.7560 |
| ChatGLM3-6b (KALLM) | Crop Variety Recommendation | 2.4000 | 2.7907 | 1.9733 | 1.3953 | 0.9867 | 1.7865 | 2.7907 | 0.0000 | 2.7907 | 0.6611 |
| | Cultivation Management | 22.4578 | 27.5028 | 22.3937 | 19.3179 | 16.9366 | 21.5378 | 25.9707 | 14.4634 | 25.9707 | 0.8293 |
| | Crop Pest and Disease Diagnosis | 0.1202 | 4.6154 | 3.2636 | 2.3077 | 1.6318 | 2.9546 | 4.7773 | 0.0000 | 4.7773 | 0.5643 |
| | Crop Pest and Disease Prevention | 3.0545 | 5.4342 | 4.2810 | 3.6305 | 3.0881 | 4.1084 | 1.9955 | 0.9710 | 1.9955 | 0.6846 |
| Llama2-7b (KALLM) | Crop Variety Recommendation | 1.6000 | 1.8605 | 1.3155 | 0.9302 | 0.6578 | 1.1910 | 1.8605 | 0.0000 | 1.8605 | 0.6665 |
| | Cultivation Management | 23.4435 | 24.2431 | 19.2610 | 16.8581 | 15.2517 | 18.9035 | 18.6848 | 12.8810 | 18.6082 | 0.8025 |
| | Crop Pest and Disease Diagnosis | 0.0015 | 0.0282 | 0.0199 | 0.0141 | 0.0100 | 0.0181 | 0.0000 | 0.0000 | 0.0000 | 0.4902 |
| | Crop Pest and Disease Prevention | 0.5141 | 0.9658 | 0.3892 | 0.2429 | 0.1667 | 0.4412 | 0.0189 | 0.0000 | 0.0189 | 0.5305 |
| Llama2-13b (KALLM) | Crop Variety Recommendation | 5.5118 | 6.5116 | 4.6044 | 3.2558 | 2.3022 | 4.1685 | 6.5116 | 0.0000 | 6.5116 | 0.6850 |
| | Cultivation Management | 26.8336 | 32.4605 | 25.7482 | 22.4743 | 19.9569 | 25.1600 | 24.9999 | 16.0181 | 24.9999 | 0.8403 |
| | Crop Pest and Disease Diagnosis | 0.0013 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0073 | 0.0000 | 0.0073 | 0.7488 |
| | Crop Pest and Disease Prevention | 1.5890 | 5.0585 | 4.2576 | 3.3273 | 2.5638 | 3.8018 | 5.1054 | 3.5714 | 5.1054 | 0.7214 |
| Qwen1.5-7b (KALLM) | Crop Variety Recommendation | 9.3023 | 10.2326 | 7.3717 | 5.2529 | 3.7299 | 6.6468 | 10.2326 | 0.4651 | 10.2326 | 0.7049 |
| | Cultivation Management | 57.8488 | 51.4228 | 45.5810 | 41.8812 | 38.9312 | 44.4541 | 49.6620 | 35.3060 | 49.6620 | 0.8729 |
| | Crop Pest and Disease Diagnosis | 50.0000 | 50.0000 | 35.3553 | 25.0000 | 17.6777 | 32.0083 | 50.0000 | 0.0000 | 50.0000 | 0.8131 |
| | Crop Pest and Disease Prevention | 14.1225 | 23.8866 | 22.3016 | 20.9342 | 19.4072 | 21.6324 | 14.2361 | 11.0294 | 14.2361 | 0.7511 |
| Qwen1.5-14b (KALLM) | Crop Variety Recommendation | 6.2745 | 7.4419 | 5.2622 | 3.7209 | 2.6311 | 4.7640 | 7.4419 | 0.0000 | 7.4419 | 0.6951 |
| | Cultivation Management | 41.1506 | 40.9353 | 34.6389 | 31.2396 | 28.6272 | 33.8602 | 36.1004 | 26.3254 | 36.1004 | 0.8759 |
| | Crop Pest and Disease Diagnosis | 0.3750 | 9.2308 | 6.5271 | 4.6154 | 3.2636 | 5.9092 | 9.2308 | 0.0000 | 9.2308 | 0.7231 |
| | Crop Pest and Disease Prevention | 3.9674 | 9.9993 | 7.6013 | 6.4210 | 5.5537 | 7.3938 | 5.5270 | 4.2640 | 5.5270 | 0.6830 |
| HuaTuo_FL_bm25 | Crop Variety Recommendation | 23.2143 | 24.0000 | 17.2635 | 12.2937 | 8.7263 | 15.5709 | 24.0000 | 1.0000 | 24.0000 | 0.7746 |
| | Cultivation Management | 0.4688 | 1.8692 | 1.6632 | 1.5659 | 1.4949 | 1.6483 | 3.3699 | 2.1990 | 3.3699 | 0.5877 |
| | Crop Pest and Disease Diagnosis | 0.0169 | 15.3846 | 10.8786 | 7.6923 | 5.4393 | 9.8487 | 15.3846 | 0.0000 | 15.3846 | 0.5750 |
| | Crop Pest and Disease Prevention | 0.3611 | 0.5416 | 0.3017 | 0.2149 | 0.1645 | 0.3057 | 0.2380 | 0.0000 | 0.2380 | 0.5292 |

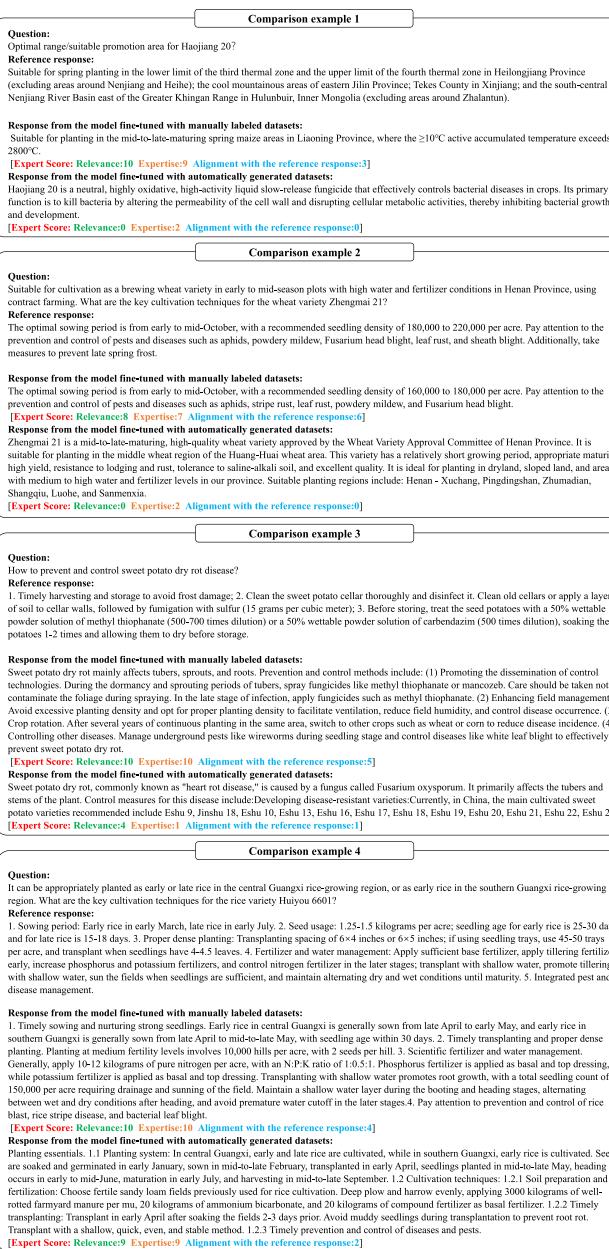


Fig. C.1. Comparison examples of model performance after fine-tuning with the manually annotated dataset and the automatically constructed dataset.

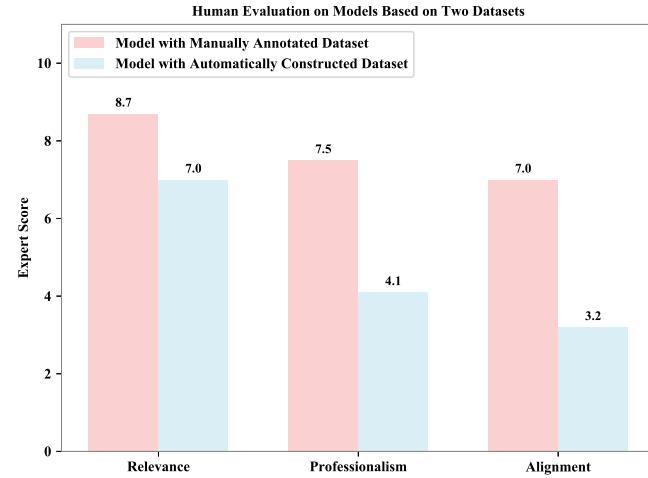


Fig. C.2. Evaluation of model responses fine-tuned with two datasets by agricultural experts. The Fleiss' Kappa is 0.84.

Query & Reference Response

Query: What are the characteristics of Jinxiangnuo No. 3?

Reference response: Jinxiangnuo No. 3 is a conventional japonica aromatic glutinous rice variety. Its full growth period is 159 days, the same as the control variety Liao Geng No. 9. The plant height is 96 cm, with a panicle length of 81.17 cm and a thousand-grain weight of 248.0 g/mu. Each panicle has a grain number of 135, with a seed-setting rate of 93.5%, a thousand-grain weight of 27.2 g/mu. The comprehensive index for rice blast disease over two years was 1.0 and 1.0, respectively, with the highest panicle blast damage rate at grade 1, indicating strong resistance to rice blast disease. The brown rice rate is 82.3%, milled rice rate is 72.9%, and head rice rate is 67.7%. The whiteness grade is 1, with a chalky rice rate of 0.5%. The alkali spreading rate is 4.7, gel consistency is 100 mm, amylose content is 2.8%, protein content is 8.46%, grain length is 5.0 mm, and the length-width ratio is 1.6, meeting the national standard for ordinary japonica glutinous rice.

KALLM (Our Model) :

This variety is a glutinous rice type. It features a compact plant type with medium height and moderate tillering ability. The panicles are relatively large with round grains that have colorless glumes and no awns. Over two years of regional trials, the average full growth period was 143.2 days, which is 0.7 days shorter than the control. The average number of effective panicles per acre was 184,000, with a panicle formation rate of 72.8%. The plant height averaged 102.9 cm, panicle length 17.8 cm, with a total grain number per panicle of 133.9, of which 111.3 were filled grains, resulting in a seed-setting rate of 83.1%. The thousand-grain weight was 23.3 grams. Resistance identification by the Provincial Academy of Agricultural Sciences and Ministry of Agriculture in 2019 showed a moderate resistance to rice blast disease, with a panicle blast damage rate of 1.0 and 1.0, respectively, and a comprehensive rice blast index of 0.8, indicating moderate resistance. The highest panicle blast damage rate was grade 7, indicating moderate susceptibility, while the highest brown pantheropper rating was grade 9, indicating high susceptibility. Quality testing by the Ministry of Agriculture Rice and Rice Products Quality Supervision, Inspection and Testing Center in 2018-2019 revealed an average head rice rate of 71.0%, a length-to-width ratio of 1.8, a gel consistency of 100.0 mm, and an amylose content of 1.4%. The overall quality indicators meet the national standards for ordinary edible rice varieties.

[Expert Score: Fluency:9 Informativeness:7 Expertise:7 Alignment with the reference response:9 Overall:8]

Nominal Pipeline:

Yield Performance of Jinxiangnuo No. 3: In 2019, it participated in the late-maturing group of special rice independent regional trials in Liaoning Province, with an average yield of 657.1 kg per mu, which was 4.7% higher than the control variety Liao Geng No. 9. In the 2020 continuation trials, the average yield was 638.4 kg per mu, 4.3% higher than the control. Over two years of regional trials, the average yield was 647.8 kg per mu, 4.5% higher than the control. In the 2021 production trials, the average yield was 646.2 kg per mu, 5.4% higher than the control variety Liao Geng No. 9.

[Expert Score: Fluency:6 Informativeness:5 Expertise:4 Alignment with the reference response:6 Overall:6]

Base:

Jinxiangnuo No. 3 is a variety of glutinous rice with the following characteristics: 1. Origin: Jinxiangnuo No. 3 is a glutinous rice variety developed by Sichuan Jinxiang Agricultural Technology Co., Ltd. 2. Approval Number: National Approval No. Yu20120013. 3. Variety Type: Short-day type, 2.2. 4. Plant Type: Compact plant type, with a semi-dwarf habit and relatively high tillering ability. 5. Panicle and Grain Structure: Each panicle contains about 120-140 grains, with a seed-setting rate of over 85%. 6. Grain Traits: The grains are yellow, semi-dentate, with a thousand-grain weight of about 250 grams and a high seed yield rate. 7. Quality Traits: Jinxiangnuo No. 3 has excellent quality, good taste, high glutinosity, and high nutritional value. 8. Resistance: It has strong resistance to stress, diseases, pests, and lodging. 9. Growth Period: In the Yangtze River Basin, the growth period of Jinxiangnuo No. 3 is 100-120 days, making it a mid-maturing variety. 10. Yield: Jinxiangnuo No. 3 has a high yield, typically reaching 500-600 kg per mu. In summary, Jinxiangnuo No. 3 is a glutinous rice variety with excellent quality, strong resistance, and high yield.

[Expert Score: Fluency:7 Informativeness:5 Expertise:5 Alignment with the reference response:5 Overall:6]

GPT3.5:

Jinxiangnuo No. 3 is a rice variety with the following main characteristics: 1. The grains are plump, the rice is translucent, and it has a fragrant, glutinous texture. 2. It has a moderate growth cycle, matures early, and possesses strong lodging resistance, making it well-suited to various climates and wide adaptability, making it suitable for cultivation in various climatic conditions. 4. The yield is stable, making it one of the high-yield rice varieties chosen for cultivation. 5. It is suitable for large-scale planting and has good market adaptability.

These are some of the characteristics of the Jinxiangnuo No. 3 rice variety, and I hope this information is helpful to you.

[Expert Score: Fluency:7 Informativeness:5 Expertise:5 Alignment with the reference response:5 Overall:6]

Fig. D.1. Responses generated by different models.

Question

What are the characteristics and features of Huiyu No. 2?

Wikipedia:

Evidence 1: Fuzhou City (Min Dong 漳州府, Pinghua: Hok-chi-ché, Min Pin: huk21-ziu5, actual pronunciation: /hūʔɿ/ tsiaɿh tsɿwɿɿ/ [Fuzhou dialect]), commonly known as Fuzhou, abbreviated as Fu or Fong, formerly known as Yecheng, and also called Mindu and Rongcheng, is the capital of Fujian Province, People's Republic of China. It is a national regional central city and one of the first coastal open cities.

Evidence 2: The Wenchuan Earthquake occurred at 14:28:04 Beijing time on May 12, 2008 (Monday). The epicenter was located near Yingxiu Town, Wenchuan County, Aba Tibetan and Qiang Autonomous Prefecture, Sichuan Province, 79 kilometers northwest of Chengdu, the provincial capital of Sichuan.

Evidence 3: Fairy Tail is a fictional organization in the Japanese manga "FAIRY TAIL," to which the main characters belong. The guild headquarters is located in Magnolia Town, on the eastern side of the Kingdom of Fiore.

Evidence 4: The Ig Nobel Prize, as the name suggests, is a parody of the Nobel Prize. It is awarded annually in early October, around the time the Nobel Prize winners are announced. It mainly honors ten achievements that "first make people laugh, and then make them think."

Evidence 5: "Cool Met Stuff" is a weather information program produced by the Hong Kong Observatory's "Weather Broadcasting Station" team since 2014, aimed at enhancing public understanding of meteorology.

Google:

Evidence 1: Online consultation, online surveys, public opinion collection, online interviews, question knowledge base, senior citizen services, convenient services, latest news, joyful travel, wonderful Guiyang, Guiyang overview, Guiyang images, Guiyang profile, Guiyang history, Guiyang tourism, government, wonderful data, latest data releases, data interpretation, data openness, statistical bulletins, statistical yearbooks, Navigation: Home > Government Affairs > Key Information Disclosure > Agriculture and Rural Affairs > Crop Management. The Guiyang Municipal Agriculture and Rural Affairs Bureau announces the 2024 Guiyang and Gui'an main promoted grain and oil crop varieties list. Date of publication: January 30, 2024, 13:37. Views: 0. Font size: [large, medium, small]. To accelerate the promotion and application of superior grain and oil crop varieties and implement the "storing grain in technology" strategy, the Guiyang Municipal Agriculture and Rural Affairs Bureau, in accordance with the requirements of the Guizhou Provincial Department of Agriculture and Rural Affairs and considering the actual conditions of Guiyang and Gui'an, has organized and selected 14 main promoted grain and oil crop varieties for 2024, which are now recommended and announced...

Evidence 2: Studies on Dry Matter Accumulation and Distributive Characteristics in Super High-yield Maize. Homepage, journal introduction, awards, editorial board, submission guidelines, advertising, subscription, contact us, English. Home > Previous issues > Volume 15, Issue 3, 2007. Previous [Next] 2007, 15(3): 095-098. Priority publication. Abstract, figures/tables, access statistics, PDF preview, references, similar literature, cited literature, attachments. [Keywords] [Abstract] Using four maize varieties as test materials, a comparative study was conducted on the characteristics of dry matter accumulation and distribution in super high-yield maize...

Evidence 3: Parental Characteristics and Seed Production Techniques of New Sunflower Hybrid 'Xin Kui 6'. Crops Journal, homepage, journal information, journal introduction, editorial board, database inclusion, publication scope, academic ethics, online journal, current issue directory, latest accepted papers, previous issues, browse abstracts, thematic abstracts, click ranking, full-text download ranking, citation ranking, author center, online submission, submission guidelines.

Evidence 4: Study on the Effect of Two New Fertilizers and Watering Irrigation Conditions on Yield and Quality Analysis of Wheat Line Shuji235 and Watering Irrigation Conditions. Agriculture - Strip Compound Planning - Water Conservancy Engineering - Wheat Fields - Facility Agriculture - Fruit Industry.

Evidence 5: Characteristics and Features of Hetian Jade. Exploring the Unique Characteristics and Features of Hetian Jade - Hetian Jade Price, Hetian Jade Price, Characteristics and Features of Hetian Jade, Exploring the Unique Characteristics and Features of Hetian Jade. Hetian Jade Price. April 13, 2024, 14:46. Characteristics and Features of Hetian Jade. Hetian jade is one of the world's famous jades with unique characteristics and features...

ChatGPT:

Evidence 1: Huiyu 2 is an intelligent conversational robot developed independently in China, with the following characteristics and features: 1. Strong natural language processing capabilities: Huiyu 2 can recognize natural language and speech, understand user intentions and information, making interactions more nuanced and natural. 2. Rich data reserves: Huiyu 2 boasts an extensive knowledge base and data reserves, providing real-time, accurate information and services across various scenarios. 3. High intelligent Q&A capability: Huiyu 2 can intelligently analyze and process user queries, offering corresponding answers and solutions through its robust logical reasoning abilities. 4. Versatile application scenarios: Huiyu 2 can be applied in multiple fields such as education, healthcare, and smart cities, offering comprehensive and convenient services to users. 5. Comprehensive security system: Huiyu 2 is equipped with a thorough information security system to ensure user privacy. It can also provide personalized services according to user needs.

Evidence 2: Huiyu 2 is a Chinese voice conversational robot with self-learning and intelligent Q&A capabilities, featuring: 1. Natural language understanding: Huiyu 2 has a powerful natural language understanding capability, able to comprehend and accurately answer most user questions. 2. Intelligent Q&A capability: Huiyu 2 has extensive knowledge reserves and deep search capabilities, quickly and accurately answering various user questions. 3. Personalized services: Huiyu 2 can provide personalized services based on user needs and preferences, such as recommending relevant information according to user interests. 4. Multi-terminal access: Huiyu 2 can be accessed through multiple terminals, such as smart speakers and mobile apps. 5. Self-learning ability: Huiyu 2 can continuously optimize its answers and service capabilities through user feedback, possessing self-learning capabilities. 6. Emotion recognition: Huiyu 2 can recognize user emotions through voice and facial expressions, providing corresponding responses and services.

Our Agricultural Knowledge Base:

Evidence 1: Yield performance of Huiyu No. 2: Participating in the 2017-2018 [Yunnan Hepo Agricultural Technology Co., Ltd. Corn New Variety Test Enterprise Consortium] mid-altitude group of ordinary corn, the average yield in the consortium trial in 2017 was 734.0 kg per mu, an increase of 14.1% compared to the control variety Haile No. 2, with a significant increase rate of 92.3%. In the 2018 continuation trial, the average yield was 857.5 kg per mu, an increase of 15.7% compared to the control variety Haile No. 2, with a significant increase rate of 100%. Over two years, the average yield in regional trials was 795.8 kg per mu, an increase of 14.9% compared to the control, with an increase rate of 96.2%. In the 2018 production test, the average yield was 790.6 kg per mu, an increase of 10.8% compared to the control, with an increase rate of 100%. It has high resistance to ear rot, resistance to northern corn leaf blight, moderate resistance to rust, and resistance to gray leaf spot and sheath blight.

Evidence 2: Review opinion: This variety meets the corn variety approval standards of Yunnan Province and is approved for cultivation. Suitable for corn planting areas at altitudes of 1000-2000 meters in Yunnan Province.

Evidence 3: Key cultivation techniques: 1. Timely sowing: According to the agricultural calendar, high-altitude spring sowings generally start around the beginning of March, while fall sowings should be limited to low-altitude summer sowing areas generally around the beginning of summer. Specific sowing times in each area can be determined according to local planting habits, climate conditions, and soil moisture or by consulting local agricultural authorities. However, spring sowing should not be earlier than March 20, and summer sowing should not be later than May 20. 2. Sowing techniques and quality requirements: a. Sowing depth: Depending on soil texture, soil moisture, and seed size, generally 4-6 cm deep, with timely soil covering after sowing. b. Fertilizer management: Before sowing, apply 1000-2000 kg of farmyard manure per mu, plus 25 kg of compound fertilizer per mu. When the plants have 5-6 leaves, apply 20 kg of urea per mu, and during the big trumpet stage, apply 25 kg of urea per mu. c. Pest and disease control: Throughout the growth period, pay attention to preventing and controlling pests and diseases. 3. Planting density: Keep 3800-4200 plants per mu. 4. Timely harvest: Harvest when the kernel milk line disappears, the husk turns yellow, and the ear is dry and loose. Pay attention to preventing and controlling gray leaf spot and sheath blight.

Evidence 4: Characteristics: The average growth period in regional trials was 133.5 days, 6.5 days later than the control. The first leaf tip of the seedling is pointed to round, with medium to strong anthocyanin coloration in the leaf sheath. The leaf blade is weakly curved, with a small angle to the stem. The leaf sheath has little or no anthocyanin coloration, and the silk has little or no anthocyanin coloration. The stem has little or no zigzag curvature. The ear has a very short shank and is conical to cylindrical, with orange-yellow intermediate-type kernels. The ear cob has little or no anthocyanin coloration. The average number of kernel rows per ear is 15.8. The 100-kernel weight is 34.6 g, with a kernel rate of 85.4%. In the 2017 consortium trial, the combined lodging rate was 1.9%, with 7.7% of test plots having a lodging rate $\geq 10.0\%$. In the 2018 consortium trial, the combined lodging rate was 0%, with 0% of test points having a lodging rate $\geq 10.0\%$. Resistance identification: high resistance to ear rot, resistance to northern corn leaf blight, moderate resistance to rust, and susceptibility to gray leaf spot and sheath blight. Quality analysis: bulk density 786 g/L, crude protein 10.8%, crude fat 5.87%, crude starch 69.1%, lysine 0.33%.

Fig. E.1. Retrieval evidence obtained from different knowledge sources.

References

[1] Y. Cao, L. Chen, Y. Yuan, G. Sun, Cucumber disease recognition with small samples using image-text-label-based multi-modal language model, Comput. Electron. Agric. 211 (2023) 107993.

Crop Variety Recommendation Example

Query:
Corn varieties suitable for planting in the Huang-Huai-Hai summer corn zone, including southern and southward areas of Henan, Shandong, Baoding and Cangzhou in Hebei, the Guanzhong Irrigation Area in Shaanxi, and parts of the plains in Yuncheng, Linfen, and Jinzhong in Shanxi?

Retrieved Knowledge:

- Jing136 approval opinion: This variety meets the national maize variety approval standards and has passed the approval. It is suitable for middle-ripening spring corn in eastern mountainous areas of Liaoning Province and parts of northern Liaoning Province, and most areas of Jilin City, Baicheng City and Tonghua City of Jilin Province.
- Characteristics of Nongda 196: Huanghuai Haixia maize group from seedling to maturity 103.0 days, the same as Zhengdan 958 control growth period.
- Yield performance of Nongda 196: Participated in the regional test of Huanghuai Haixia corn Group in 2019-2020, and the average yield per mu in two years was 671.8 kg, which was 4.0% higher than that of Zhengdan 958.
- L1748 Certification opinion: The variety meets the national corn variety certification standards and passed the certification. It is suitable for planting in Henan Province, Shandong Province, south and south areas of Baoding City and Cangzhou City in Hebei Province, Guanzhong Irrigation Area in Shaanxi Province.....
- Nongshen 196 certification opinions: This variety meets the national corn variety certification standards and has passed the certification. It is suitable for Huanghuai summer corn area in Henan Province, Shandong Province, Baoding City and Cangzhou City in Hebei Province, and Guanzhong irrigation area in Shaanxi Province.....
- Nongshen 196 cultivation technology key points: 1. Sowing period. Sow in early to mid-June. 2. Sowing method. Mechanical drill sowing, precision sowing.....

Fig. F.1. The query and retrieved knowledge of Crop Variety Recommendation example.

[2] M.T. Kuska, M. Wahabzada, S. Paulus, AI for crop production—Where can large language models (LLMs) provide substantial value? Comput. Electron. Agric. 221 (2024) 108924.

[3] J. Qing, X. Deng, Y. Lan, Z. Li, GPT-aided diagnosis on agricultural image based on a new light YOLOPC, Comput. Electron. Agric. 213 (2023) 108168.

[4] J. Li, M. Xu, L. Xiang, D. Chen, W. Zhuang, X. Yin, Z. Li, Foundation models in smart agriculture: Basics, opportunities, and challenges, Comput. Electron. Agric. 222 (2024) 109032.

[5] R. Shen, M. Li, C. Zhao, B. Wang, Y. Guan, J. Liu, J. Jiang, Hierarchical causal discovery from large-scale observed variables, IEEE Transactions on Knowledge and Data Engineering (2025) 1–14.

[6] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Adv. Neural Inf. Process. Syst. 35 (2022) 27730–27744.

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

[8] H. Touvron, T. Lavril, G. Izard, X. Martinet, M.A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, 2023, arXiv preprint arXiv:2302.13971.

[9] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Tang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, et al., Qwen technical report, 2023, arXiv preprint arXiv:2309.16609.

[10] H. Wang, X. Du, W. Yu, Q. Chen, K. Zhu, Z. Chu, L. Yan, Y. Guan, Apollo's oracle: retrieval-augmented reasoning in multi-agent debates, 2023, arXiv preprint arXiv:2312.04854.

[11] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, et al., A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023, arXiv preprint arXiv:2311.05232.

[12] B. Silva, L. Nunes, R. Estevão, R. Chandra, GPT-4 as an agronomist assistant? Answering agriculture exams using large language models, 2023, arXiv preprint arXiv:2310.06225.

[13] A. Gupta, A. Shirgaonkar, A.d. Balaguer, B. Silva, D. Holstein, D. Li, J. Marsman, L.O. Nunes, M. Rouzbahman, M. Sharp, et al., RAG vs fine-tuning: pipelines, tradeoffs, and a case study on agriculture, 2024, arXiv preprint arXiv:2401.08406.

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

[15] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, et al., Constitutional ai: Harmlessness from ai feedback, 2022, arXiv preprint arXiv:2212.08073.

[16] T. Wei, L. Zhao, L. Zhang, B. Zhu, L. Wang, H. Yang, B. Li, C. Cheng, W. Lü, R. Hu, et al., Skywork: A more open bilingual foundation model, 2023, arXiv preprint arXiv:2310.19341.

[17] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, Y. Zhang, Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge, Cureus 15 (6) (2023).

[18] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, 2021, arXiv preprint arXiv:2106.09685.

[19] Z. Bao, W. Chen, S. Xiao, K. Ren, J. Wu, C. Zhong, J. Peng, X. Huang, Z. Wei, Disc-medilm: Bridging general large language models and real-world medical consultation, 2023, arXiv preprint arXiv:2308.14346.

[20] H. Wang, C. Liu, N. Xi, Z. Qiang, S. Zhao, B. Qin, T. Liu, Huatuo: Tuning llama model with chinese medical knowledge, 2023, arXiv preprint arXiv:2304.06975.

[21] H. Zhang, J. Chen, F. Jiang, F. Yu, Z. Chen, G. Chen, J. Li, X. Wu, Z. Zhiyi, Q. Xiao, et al., HuatuoGPT, towards taming language model to be a doctor, in: Findings of the Association for Computational Linguistics, EMNLP 2023, 2023, pp. 10859–10885.

18

- [22] Q. Huang, M. Tao, C. Zhang, Z. An, C. Jiang, Z. Chen, Z. Wu, Y. Feng, Lawyer llama technical report, 2023, arXiv preprint [arXiv:2305.15062](#).
- [23] J. Cui, Z. Li, Y. Yan, B. Chen, L. Yuan, Chatlaw: Open-source legal large language model with integrated external knowledge bases, 2023, arXiv preprint [arXiv:2306.16092](#).
- [24] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, Z. Sui, A survey on in-context learning, 2022, arXiv preprint [arXiv:2301.00234](#).
- [25] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, Retrieval-augmented generation for large language models: A survey, 2023, arXiv preprint [arXiv:2312.10997](#).
- [26] Y. Deng, X. Zhang, H.Y. Huang, Y. Hu, Towards faithful dialogues via focus learning, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 4554–4566.
- [27] M. Mei, Y. Hu, Y. Deng, X. Zhang, Y. Li, H. You, Summarizing community-based question-answer pairs with focus rectification, in: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2024, pp. 11391–11395.
- [28] O. Khattab, M. Zaharia, Colbert: Efficient and effective passage search via contextualized late interaction over bert, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 39–48.
- [29] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, *Adv. Neural Inf. Process. Syst.* 33 (2020) 9459–9474.
- [30] R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders, et al., Webgpt: Browser-assisted question-answering with human feedback, 2021, arXiv preprint [arXiv:2112.09332](#).
- [31] Y. Qin, Z. Cai, D. Jin, L. Yan, S. Liang, K. Zhu, Y. Lin, X. Han, N. Ding, H. Wang, et al., WebCPM: interactive web search for Chinese long-form question answering, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 8968–8988.
- [32] Z. Jiang, F.F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, J. Callan, G. Neubig, Active retrieval augmented generation, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 7969–7992.
- [33] A. Asai, Z. Wu, Y. Wang, A. Sil, H. Hajishirzi, Self-RAG: learning to retrieve, generate, and critique through self-reflection, in: The Twelfth International Conference on Learning Representations, 2023.
- [34] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.T. Yih, Dense passage retrieval for open-domain question answering, in: 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Association for Computational Linguistics (ACL), 2020, pp. 6769–6781.
- [35] J. Ni, C. Qu, J. Lu, Z. Dai, G.H. Abrego, J. Ma, V. Zhao, Y. Luan, K. Hall, M.W. Chang, et al., Large dual encoders are generalizable retrievers, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 9844–9855.
- [36] A. Askari, A. Abolghasemi, G. Pasi, W. Kraaij, S. Verberne, Injecting the BM25 score as text improves BERT-based re-rankers, in: European Conference on Information Retrieval, Springer, 2023, pp. 66–83.
- [37] C. Tang, H. Zhang, T. Loakman, C. Lin, F. Guerin, Terminology-aware medical dialogue generation, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2023, pp. 1–5.
- [38] N. Reimers, I. Gurevych, Sentence-BERT: sentence embeddings using siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019.
- [39] M. Kale, A. Rastogi, Text-to-text pre-training for data-to-text tasks, in: Proceedings of the 13th International Conference on Natural Language Generation, 2020, pp. 97–102.
- [40] K. Lv, Y. Yang, T. Liu, Q. Gao, Q. Guo, X. Qiu, Full parameter fine-tuning for large language models with limited resources, 2023, arXiv preprint [arXiv:2306.09782](#).
- [41] C.Y. Chuang, J. Robinson, Y.C. Lin, A. Torralba, S. Jegelka, Debiased contrastive learning, *Adv. Neural Inf. Process. Syst.* 33 (2020) 8765–8775.
- [42] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, J. Tang, GLM: general language model pretraining with autoregressive blank infilling, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 320–335.
- [43] J.L. Fleiss, B. Levin, M.C. Paik, et al., The measurement of interrater agreement, *Stat. Methods Rates Proportions* 2 (212–236) (1981) 22–23.
- [44] J. Kaplan, S. McCandlish, T. Henighan, T.B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling laws for neural language models, 2020, arXiv preprint [arXiv:2001.08361](#).