applied sciences

Article

Enhancing Plant Protection Knowledge with Large Language Models: A Fine-Tuned Question-Answering System Using LoRA

Jie Xiong, Lingmin Pan, Yanjiao Liu, Lei Zhu, Lizhuo Zhang and Siqiao Tan *

College of Information and Intelligence, Hunan Agricultural University, Changsha 410128, China;
xj990328@stu.hunau.edu.cn (J.X.); lingminpan@stu.hunau.edu.cn (L.P.); liuyj213@stu.hunau.edu.cn (Y.L.);
leizhu@hunau.edu.cn (L.Z.); zhanglizhuo@hunau.edu.cn (L.Z.)
* Correspondence: tsq@hunau.edu.cn

Abstract: To enhance the accessibility and accuracy of plant protection knowledge for agricultural practitioners, this study develops an intelligent question-answering (QA) system based on a large language model (LLM). A local knowledge base was constructed by vectorizing 7000 research papers and books in the field of plant protection, from which 568 representative papers were selected to generate QA data using an LLM. After data cleaning and filtering, a fine-tuning dataset comprising 9000 question–answer pairs was curated. To optimize the model's performance, low-rank adaptation (LoRA) was applied to the InterLM-20B model, resulting in the InterLM-20B-LoRA, which was integrated with Langchain-ChatChat and the local knowledge base to develop the QA system. Additionally, retrieval-augmented generation (RAG) technology was implemented to enhance response accuracy by enabling the model to retrieve relevant field-specific knowledge before generating answers, effectively mitigating the risk of hallucinated information. The experimental results demonstrate that the proposed system achieves an answer accuracy of 97%, highlighting its potential as an advanced solution for intelligent agricultural QA services.

Keywords: plant protection; QA system; LLMs; low-rank adaptive; retrieval-augmented generation

Academic Editor: Andrea Prati

Received: 28 February 2025
Revised: 27 March 2025
Accepted: 27 March 2025
Published: 1 April 2025

Citation: Xiong, J.; Pan, L.; Liu, Y.; Zhu, L.; Zhang, L.; Tan, S. Enhancing Plant Protection Knowledge with Large Language Models: A Fine-Tuned Question-Answering System Using LoRA. *Appl. Sci.* **2025**, *15*, 3850. https://doi.org/10.3390/app15073850

1. Introduction

Plant protection is a critical component of modern agricultural production, encompassing pest and disease management, soil health preservation, and crop yield optimization [1,2]. The rapid advancement of agricultural practices has introduced complex challenges, including the emergence of novel plant pathogens, increasing pest resistance to chemical control measures, and the impacts of climate change on crop productivity [3,4]. Accurate, timely plant protection knowledge is essential. This is particularly true for farmers, who are the primary field practitioners and often lack specialized expertise. We aim to provide them with practical, real-time recommendations. These guidelines are designed to address their daily challenges effectively. However, conventional knowledge acquisition methods, such as expert consultations, printed reference materials, and fragmented online resources, are often time-consuming, require specialized expertise, and may not provide real-time, context-specific recommendations. Moreover, variations in crop species, regional environmental conditions, and pest dynamics necessitate tailored solutions, making generic knowledge dissemination approaches insufficient to address the diverse and evolving needs of the agricultural sector.

In recent years, large language models (LLMs) have demonstrated exceptional natural language understanding and generation capabilities, finding applications in diverse fields such as healthcare, finance, and legal services [5–7]. Their ability to process complex queries and generate contextually relevant responses makes them promising candidates for question-answering (QA) systems. However, applying LLMs to highly specialized domains such as plant protection presents significant challenges. One major limitation is their lack of domain-specific knowledge, as pre-trained models rely primarily on publicly available datasets that often lack high-quality, expert-reviewed agricultural content. Consequently, these models may generate incomplete, outdated, or misleading responses when addressing plant protection issues. Additionally, LLMs are prone to the well-documented hallucination problem, where they produce factually incorrect or unverifiable information. In plant protection, where accurate and scientifically validated knowledge is essential for effective decision-making, such hallucinations can result in incorrect diagnoses, improper pesticide applications, and ineffective pest management strategies, potentially leading to economic losses and environmental harm [8]. Furthermore, traditional LLMs struggle to adapt to region-specific agricultural conditions, as factors such as climatic variations, soil characteristics, and pest dynamics necessitate dynamic and context-aware responses that generic models fail to provide.

To mitigate these challenges, fine-tuning has been widely explored to enhance LLM performance in specialized domains. However, conventional fine-tuning modifies a large number of parameters, requiring extensive computational resources and substantial storage, making it impractical for organizations with limited hardware capabilities. A more efficient alternative is low-rank adaptation (LoRA), a parameter-efficient fine-tuning method that significantly reduces computational overhead while preserving model performance. LoRA achieves this by introducing trainable low-rank matrices within specific layers of the pre-trained model instead of modifying the full parameter set, enabling efficient adaptation to new tasks [9,10]. This approach not only enhances accessibility for domain-specific applications but also ensures that the model retains its core linguistic capabilities. In this study, we leverage LoRA to fine-tune an LLM for plant protection, equipping it with specialized knowledge while maintaining computational efficiency.

In addition to fine-tuning, retrieval-augmented generation (RAG) has emerged as a powerful technique for enhancing the accuracy and reliability of LLM responses. RAG integrates information retrieval with text generation, allowing the model to access an external knowledge base when formulating responses dynamically. Unlike traditional LLMs that rely solely on static training data, RAG-based models can retrieve relevant information from a curated document corpus, thereby reducing hallucinations and ensuring that responses are grounded in authoritative sources [11]. In this study, we construct a local knowledge base consisting of professional scientific literature on plant protection, which serves as the retrieval component of our RAG QA system. This enables the model to reference validated agricultural research and expert-authored content when answering user queries, improving response quality and trustworthiness. The main contributions of this article mainly include the following points:

- We develop an intelligent QA system tailored for the plant protection domain, addressing the critical need for precise and accessible agricultural knowledge. By leveraging natural language processing (NLP), we bridge the gap between complex scientific research and practical decision-making, enabling farmers and agricultural practitioners to obtain expert-level insights efficiently. The system is designed to accommodate diverse user needs, providing customized recommendations based on varying geographical conditions and crop-specific requirements.
- We integrate advanced fine-tuning and retrieval strategies to optimize LLM performance for plant protection applications. By combining LoRA-based fine-tuning with RAG-enhanced retrieval, we mitigate common limitations of pre-trained models, such

as hallucinations and domain knowledge deficiencies. Additionally, we construct a local domain-specific knowledge base composed of peer-reviewed scientific publications, ensuring that the model's responses are accurate, authoritative, and explainable. This hybrid approach significantly enhances the reliability of LLM-generated answers in an agricultural context.

- We contribute a high-quality plant protection dataset containing over 9000 question–answer pairs, covering fundamental plant protection principles and expert-verified knowledge extracted from academic papers. This dataset serves as a valuable resource for training and evaluating domain-adapted language models. By fine-tuning our model on this dataset using LoRA, we achieve substantial improvements in domain-specific comprehension and response accuracy, demonstrating the effectiveness of our approach in equipping LLMs with specialized agricultural knowledge.

2. Related Work

The development of question-answering systems can be traced back to the 1960s when early systems were primarily designed for specific domains and structured data [12]. For instance, in 1961, Green developed the BASEBALL [13] system to answer questions about a single season of Major League Baseball; the LUNAR [14] system was created to provide geological analyses of lunar rock samples from the Apollo moon landing program. These pioneering systems relied on early artificial intelligence techniques to deliver accurate answers within narrowly defined fields.

In the 1970s, advances in computational linguistics shifted the research focus toward employing linguistic technologies to reduce the costs of system construction. Although these systems remained confined to specific domains—for example, the Unix Consultant [15] system, which answered questions about the Unix operating system using a manually designed domain knowledge base—these early efforts laid a solid foundation for the evolution of question-answering technologies.

The 1990s marked a significant expansion as question-answering systems began to tackle open-domain challenges. In 1999, the introduction of the TREC-8 (Text Retrieval Conference) question-answering competition heralded the birth of open-domain systems capable of addressing a broader array of topics, thus breaking free from the constraints of earlier, domain-specific approaches. Entering the 21st century, IBM's Watson [16] system famously defeated human contestants on the television quiz show "Jeopardy!", vividly demonstrating the potent capabilities of open-domain question-answering systems.

More recently, the advent of deep learning has further propelled the development of question-answering systems. Models built on the Transformer architecture, such as BERT [17] and GPT-3 [18], have showcased outstanding natural language understanding and generation abilities through pre-training on vast corpora. These models not only manage complex language phenomena but have also spurred the rise in generative artificial intelligence, leading to systems like ChatGPT and LLaMA. These advanced models generate fluent, natural content and are adept at addressing intricate, cross-domain challenges and finding applications in customer service, education, and content creation.

Driven by a rich historical evolution, new applications tailored to specific fields continue to emerge, for example, FarmChat, developed by Mohon Jain et al. [19]. In 2018, it is a dialog agent that supports both text and voice inputs by leveraging call center logs from the Kisan system and insights from local agricultural experts to deliver agricultural information to rural Indian farmers. Similarly, the crop protection information system proposed by Tende et al. [20] employs SMS and web to disseminate plant protection advice to Tanzanian rural farmers, enabling timely responses to pests and diseases. In 2024, Koopman et al. [21] introduced AgAsk, an intelligent agent that retrieves scientific literature to provide farmers with reliable agricultural information, thereby aiding in informed planting decisions, while the Template-based Chatbot for Agriculture Related FAQs by Zhang et al. [22] utilizes predefined templates and rules to rapidly generate answers, significantly enhancing the efficiency of information retrieval. Dhavale et al. [23] integrated generative adversarial networks for image enhancement, convolutional neural networks for precise disease detection, and large language models for interactive farmer support, thereby revolutionizing crop monitoring and management. Similarly, Klair et al. [24] demonstrated that generative AI techniques can offer accurate and scalable diagnostic solutions for plant diseases. Moreover, Madaan et al. [25] proposed a framework that combines natural language processing with conventional machine learning to deliver context-aware and tailored disease detection. In addition, Majumder and Khandelwal [26] explored the use of computer vision and generative AI to forecast crop yields and optimize resource planning, while Fahim-Ul-Islam et al. [27] presented a robust framework leveraging transformer models and federated learning to identify wheat leaf diseases in distributed agricultural environments accurately.

In the medical field, analogous technical innovations have been realized. DoctorGPT, proposed by Li et al. [28] in 2023, is a large language model designed for Chinese medical question answering. Based on the open-source Baichuan2 model and fine-tuned with LoRA technology, DoctorGPT met the specific demands of the medical domain, with experimental results demonstrating its ability to provide accurate and professional advice. Likewise, DoctorGLM, introduced by Xiong et al. [29] in 2023, employs LoRA to fine-tune a pre-trained model and incorporates a custom-built medical QA dataset, markedly enhancing its performance in Chinese medical question-answering tasks.

3. Materials and Methods

3.1. Plant Protection Dataset

In this study, we constructed a high-quality, comprehensive plant protection dataset comprising 7000 professional papers and books [30]. During the data screening process, to ensure the quality and consistency of the QA data, we de-weighted the data, standardized the format, and filtered the low-quality entries. First, among 7000 agriculture-related documents, we performed de-weighting using DOI, title similarity, and abstract matching to eliminate duplicates or highly similar documents and used MinHash to calculate text similarity to remove redundant or similarly expressed QA pairs to ensure the diversity and independence of the data. In the process of format standardization, we conducted text cleanup, including the removal of HTML tags, and special characters, the unification of full-angle and half-angle symbols, and the standardization of the use of punctuation. In addition, we standardized the sentence format to maintain the consistency of the question, for example, “How to control rice diseases?” to “What are the methods of controlling rice diseases?”, and ensured the consistency of dosage expressions, such as “2 g/L” to “2 g/L”. In addition, to reduce colloquial expressions, we changed “dosing” to “pesticide spraying” to standardize the language. In terms of low-quality data filtering, we used Perplexity (PPL) to detect language fluency and excluded pairs with high PPL values (>100) to ensure the readability of the text. At the same time, we filtered QA with answer lengths of less than 10 words or more than 500 words to avoid content with too little information or too long redundancy. Through the above data cleaning and quality control, we finally constructed 9000 high-quality agricultural QA datasets, which provide reliable data support for fine-tuning the QA system [31]. Experimental analysis demonstrates that this dataset significantly enhances the F1 score of our QA system. It aggregates papers from multiple academic databases and professional books from agricultural research institution libraries, covering core areas such as pest and disease management, crop pathology, and pesticide application. These materials offer not only rich theoretical knowledge but also experimental data and practical application cases, thereby providing a robust scientific foundation for the system's knowledge base.

To further broaden the dataset and enhance the model's generalization, we employed large language model to generate a series of question–answer pairs using designed prompt templates [32,33]. These templates address fundamental issues in plant protection, including crop disease prevention and treatment, as well as strategies to optimize the growth environment and reduce disease incidence. The generated question-answer pairs were rigorously screened and cleaned to ensure scientific validity and practical relevance, effectively supplementing gaps in the original dataset and deepening the model's understanding of core issues. All generated pairs underwent manual refinement to ensure high quality and reliability, thereby enhancing the model's capacity to address fundamental questions and bolstering its generalization in real-world applications.

Since the original dataset comprised files and text in various formats, we implemented a series of preprocessing steps to ensure efficient processing and learning. These steps involved clearing sensitive information and redundant content, removing low-quality data, performing deduplication to ensure dataset independence, and reformatting the data into the question–answer format required for training. Additionally, we built a local vector database to convert the processed data into vector embeddings, facilitating efficient retrieval during answer generation.

Through meticulous preprocessing and data integration, we successfully created a plant protection question-answering dataset comprising over 9000 question–answer pairs and a local knowledge base containing 7000 professional papers and books, thereby providing robust data support for our system.

3.2. Methods

3.2.1. InternLM Model

InternLM [34] was developed through a collaborative effort by the Shanghai Artificial Intelligence Laboratory, SenseTime, the Chinese University of Hong Kong, Fudan University, and Shanghai Jiao Tong University. Built upon the Transformer architecture, the model is designed to tackle complex multi-task natural language processing (NLP) challenges. Pre-trained on diverse high-quality datasets, InternLM further enhances its language understanding and generation capabilities through self-supervised learning. The model is optimized for multilingual scenarios. It not only processes Chinese fluently but also exhibits strong cross-lingual capabilities, particularly in applications such as dialog generation and question-answering systems. Leveraging its large-scale parameters, InternLM exhibits strong adaptability and efficiency in complex reasoning tasks.

In terms of architecture optimization, InternLM integrates a variety of innovative technologies to improve the training and reasoning efficiency of the model. First, during training, the model employs a hybrid parallelization strategy that integrates data parallelism, model parallelism, and pipeline parallelism, significantly enhancing distributed training efficiency. In addition, by incorporating a sparse attention mechanism, InternLM effectively reduces computational overhead in processing long text sequences, thereby enhancing inference speed. Notably, the model supports multimodal data fusion, enabling comprehensive analysis of text, image, and audio information while enhancing performance in cross-modal tasks. Finally, self-supervised learning enables the model to train on vast amounts of unlabeled data, further enhancing its generalization ability, particularly in multilingual tasks.

3.2.2. Low-Rank Adaptation

Low-rank adaptation (LoRA) [35] is an advanced fine-tuning strategy specifically designed to efficiently optimize the parameters of large pre-trained language models. Instead of modifying all parameters, LoRA optimizes model weights by introducing low-rank matrices, which dramatically reduces the number of trainable parameters during fine-tuning. Compared to traditional full-parameter fine-tuning methods, LoRA not only lowers computational costs but also preserves the model's core parameters, thereby mitigating the risk of overfitting. Its ability to quickly adapt to new tasks while maintaining overall model stability makes LoRA particularly well-suited for domain-specific applications that require frequent adaptation.

In traditional full-parameter fine-tuning, all model parameters must be updated for a new task. In contrast, LoRA significantly reduces the number of trainable parameters by decomposing the weight matrix into two low-rank matrices. LoRA introduces two low-rank matrices, A and B, with smaller dimensions to approximate the original weight matrix. During fine-tuning, only these low-rank matrices are updated, leaving the original weight matrix unchanged.

Assuming that the weight matrix of a large language model is W_0, LoRA updates it in the following way:

$$h = W_0 x + W x = W_0 x + BAx \tag{1}$$

where $W = BA$. Due to the low rank of A and B, the number of updated parameters is drastically reduced. In this way, the computational complexity and storage requirements of the model are also reduced.

In this study, we applied LoRA to enhance the performance of large language models in plant protection question-answering tasks. Compared to traditional full-parameter fine-tuning, LoRA updates only the low-rank matrices of specific layers, reducing computational overhead while preserving the model's original knowledge structure. By applying LoRA to the InternLM-20B model, we improved its ability to model professional terminology and domain-specific knowledge in plant protection.

3.2.3. Retrieval-Augmented Generation

Retrieval-augmented generation [36,37] (RAG) leverages the strengths of both retrieval and generation models to enhance the performance of large language models in knowledge-intensive tasks. While traditional generative models can generate fluent text, they often produce misleading or incomplete responses when handling queries requiring specific domain knowledge due to a lack of accurate contextual information. In contrast, RAG incorporates retrieval mechanisms to effectively leverage local knowledge bases during answer generation, thereby producing more accurate and well-supported responses.

The RAG framework comprises two key components: the retrieval component and the generation component.

Retrieval Component: Upon receiving a user query, the system employs a predefined retrieval model to search the local knowledge base for the most relevant documents. The retrieved documents are transformed into vector representations using embedding techniques, enabling efficient similarity searches.

Generation Component: Once relevant documents are retrieved, the generation component processes them. It typically consists of a large pre-trained language model that integrates the retrieved documents with the user query to generate responses. This mechanism enables the model to generate responses based on the latest and most relevant information from the knowledge base, thereby enhancing accuracy and comprehensiveness.

In the RAG framework, the generation component does not solely rely on the model's internal knowledge but dynamically integrates information from the local knowledge base to generate responses. This approach significantly reduces the 'hallucination' phenomenon commonly observed in traditional generative models when addressing domain-specific queries, reducing the risk of generating inaccurate or misleading information.

This study develops a plant protection knowledge question-answering system leveraging RAG technology to enhance answer accuracy and reliability. The system integrates the generative capabilities of a fine-tuned large language model with dynamic retrieval of professional literature, research papers, and technical guidelines in plant protection, ensuring responses that are timely, relevant, and domain-specific.

The FAISS vector library is employed to convert knowledge base documents into vector representations. When a user submits a query, it is transformed into a vector and matched against the knowledge base using similarity search, enabling rapid retrieval of the most relevant information. The retrieved documents serve as contextual information, which, along with the user query, is fed into the generation module to produce responses that incorporate expert plant protection knowledge and accurately address complex plant pest and disease issues.

The integration of RAG technology significantly enhances the system's capability to handle multi-domain knowledge and complex agricultural scenarios, delivering more precise and contextually relevant responses to users.

3.2.4. Evaluation Metrics

The evaluation of large language models covers multiple dimensions [38], including language generation ability, task performance, comprehension depth, and logical reasoning. Automatic evaluation employs computational metrics to quantitatively assess both the quality of generated text and overall task performance. Common evaluation metrics include BLEU [39], ROUGE [40], BERTScore [41], QAGS [42], FactCC [43], and FEQA [44], which measure aspects such as text similarity, semantic consistency, and fluency. However, BLEU rely on precise n-gram matching and are less sensitive to semantic equivalence and syntactic variations. Furthermore, although QAGS and FEQA can indirectly measure text fidelity through question–answer generation, their question-generation process often introduces significant uncertainty, and their correlation with human judgment remains limited in practical applications. For these reasons, we have chosen not to use BLEU, QAGS, and FEQA, and instead focus on metrics that more accurately capture semantic and factual consistency.

In this study, we employ ROUGE, BERTScore, and FactCC as automatic evaluation metrics to assess model performance. These three methods are both efficient and objective, providing a comprehensive measure of the model's performance on generation tasks. ROUGE primarily measures the overlap between generated and reference text, making it particularly suitable for assessing summarization accuracy and other text generation tasks. BERTScore leverages a deep language model to embed both the generated text and the reference text, then calculates the cosine similarity between these embeddings to capture semantic information. FactCC evaluates factual consistency by using a supervised binary classifier to determine whether the factual statements in the generated text align with authoritative data, effectively identifying hallucinations and factual errors. In summary, these indicators effectively capture text quality and semantic accuracy while reducing the subjectivity and time costs associated with manual evaluation, thereby providing a more reliable basis for the performance evaluation of large-scale language models.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is primarily used to measure the similarity between generated and reference texts based on *N*-gram overlap. An *N*-gram is a contiguous sequence of *N* items—such as characters, syllables, or words—in a text, where *N* is a positive integer denoting the sequence length. ROUGE is widely applied in automatic summarization and machine translation tasks. Common ROUGE variants include *ROUGE-N* and *ROUGE-L*.

ROUGE-N quantifies the *N*-gram overlap between generated and reference texts. The equation is:

$$ROUGE - N = \frac{\sum\limits_{S\in\{ReferenceSummaries\}} \sum\limits_{gram_N\in S} Count_{match}(gram_N)}{\sum\limits_{S\in\{ReferenceSummaries\}} \sum\limits_{gram_N\in S} Count(gram_N)} \tag{2}$$

The numerator of the equation calculates the minimum number of matches of all *N*-grams between the generated text and the reference text, while the denominator is the total number of *N*-grams in the reference text.

ROUGE-L is based on the longest common subsequence (LCS) to capture the structural similarity between the generated text and the reference text. The equation is expressed as follows:

$$ROUGE - L = \frac{LCS(gen, ref)}{length(ref)} \tag{3}$$

where $LCS(gen, ref)$ is the length of the longest common subsequence (LCS) between the generated text and the reference text, and $Length(ref)$ is the length of the reference text; ROUGE-L is mainly used to calculate the similarity between the generated text and the reference text in terms of sentence structure, which is applicable to the scenario of structured information.

BERTScore is a metric designed to measure text similarity. It evaluates the similarity between two sentences by calculating the cosine similarity between the embedding vectors produced by a pre-trained BERT model.

The core idea behind BERTScore is to input two sentences individually into the BERT model, obtain their vector representations, and then compare these vectors. Unlike traditional similarity methods based on static word vectors, BERTScore more accurately assesses sentence similarity by incorporating contextual information and semantic features. In particular, for sentences that convey similar meanings despite differing in expression, BERTScore more effectively captures their underlying similarity.

$$R_{BERT} = \frac{1}{|X|} \sum_{X_i\in X} \max_{S_j\in S} X_i^T S_j \tag{4}$$

$$P_{BERT} = \frac{1}{|S|} \sum_{S_j\in S} \max_{X_i\in X} X_i^T S_j \tag{5}$$

$$F_{BERT} = 2\frac{P_{BERT} \times R_{BERT}}{P_{BERT} + R_{BERT}} \tag{6}$$

where X denotes the reference text, X_i denotes the i-th reference text, S denotes the generated text, S_j denotes the j-th generated text, and X_i^T denotes the transpose matrix.

FactCC is a metric designed to measure factual consistency by determining whether the generated text is consistent with the reference text. It relies solely on the reference text to verify factual alignment. The core idea behind FactCC is to input each generated sentence along with the reference text into a supervised classifier that has been trained on labeled data, indicating whether a sentence is factually consistent. By averaging the classifier's outputs of overall generated sentences, FactCC produces an overall factual consistency score. The calculation is as follows:

$$C_{\text{FactCC}}(S, R) = \frac{1}{|S|} \sum_{s \in S} P(C = 1 \mid s, R) \tag{7}$$

where $C_{\text{FactCC}}(S, R)$ denotes the factual consistency score for the generated text S given the reference text R, and $P(C = 1 \mid s, R)$ represents the classifier's probability that the generated sentence s is factually consistent with R.

4. Experimental Results and Analyses

4.1. Experimental Setup and Evaluation Metrics

The experimental environment, as summarized in Table 1, was established on a robust Linux-based system running Linux-4.19.36 + x86_64 with glibc2.28, leveraging the computational power of a Huawei Ascend 910B NPU. The implementation was carried out using Python 3.9.18, and our framework integrated LangChain-ChatChat alongside Fastchat to facilitate interactive functionalities. For text segmentation, we employed the ChineseRecursiveTextSplitter. Our system utilized the Internlm-20B-LoRA as the large language model, with the ms-bge@ascend serving as the embedding model. This configuration provided a stable and efficient platform for evaluating the performance of our intelligent question-answering system in the agricultural domain.

Table 1. Configuration information of experimental environment.

Item	Configuration
OS	Linux-4.19.36 + -x86_64-with-glibc2.28
NPU	Huawei ascend 910B
Programming Language	Python 3.9.18
LangChain-ChatChat version	0.2.8
Fastchat version	0.2.33
Splitter	ChineseRecursiveTextSplitter
LLM	Internlm-20B-LoRA
Embeddings model	ms-bge@ascend

The experimental environment, as summarized in Table 1, was established on a robust Linux-based system running Linux-4.19.36+ x86_64 with glibc2.28, leveraging the computational power of a Huawei Ascend 910B NPU. The implementation was carried out using Python 3.9.18, and our framework integrated LangChain-ChatChat alongside Fastchat to facilitate interactive functionalities. For text segmentation, we employed the ChineseRecursiveTextSplitter. Our system utilized the Internlm-20B-LoRA as the large language model, with the ms-bge@ascend serving as the embedding model. This configuration provided a stable and efficient platform for evaluating the performance of our intelligent question-answering system in the agricultural domain.

This paper uses ROUGE, BERTScore, and FactCC to evaluate model performance. ROUGE (calculated using Equations (2) and (3)) measures the n-gram overlap between generated and reference answers, while BERTScore (computed via Equations (4)–(6)) assesses semantic similarity using a pre-trained BERT model to gauge answer accuracy and fluency. FactCC evaluates factual consistency by using a supervised binary classifier to determine whether the factual statements in the generated text align with authoritative data. Table 2 presents the performance of various models according to these metrics.

Table 2. Experimental results of the plant protection QA system and other reference models on the plant protection dataset.

Models	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**BERTScore**	**FactCC**
				F1 Score	**F1 Score**
InternLM-20B	0.324	0.121	0.200	0.936	0.829
InternLM-20B-LoRA	0.394	0.139	0.254	0.953	0.846
BaiChuan-13B	0.317	0.114	0.189	0.938	0.832
ChatGPT	0.364	0.159	0.239	0.945	0.857
Plant Protection QA System	**0.578**	**0.321**	**0.377**	**0.970**	**0.897**

Note: the best results are highlighted in bold.

4.2. Comparison with Other Models

This paper compares the performance of several models in plant protection question-answering tasks, including InternLM-20B, InternLM-20B-LoRA, BaiChuan-13B, ChatGPT, and our dedicated plant protection question-answering system. As shown in Table 2, based on ROUGE, BERTScore, and FactCC metrics, our system exhibits significant advantages. Specifically, the ROUGE-1 and ROUGE-L scores of 0.578 and 0.377, respectively, indicate that the generated text closely matches the reference answers in terms of keyword coverage and overall expression. Moreover, a ROUGE-2 score of 0.321 confirms its superior phrase-level accuracy, demonstrating a more precise handling of complex sentences in the domain.

BERTScore evaluation indicates that our plant protection question-answering system achieved F1 scores of 0.970, demonstrating superior semantic understanding and generation accuracy. In contrast, the InternLM-20B-LoRA model—fine-tuned with LoRA—recorded F1 scores of 0.953, underscoring the effectiveness of fine-tuning in enhancing domain adaptability. General-purpose models such as ChatGPT and BaiChuan-13B performed relatively poorly; although ChatGPT shows competitiveness in semantic evaluation, its accuracy and domain-specific professionalism remain inferior to those of models optimized for plant protection.

InternLM-20B achieves a FactCC score of 0.829, while InternLM-20B-LoRA—fine-tuned with LoRA—improves this score to 0.846, indicating that meticulous fine-tuning plays a positive role in enhancing domain adaptability and factual consistency. BaiChuan-13B scores 0.832, similar to InternLM-20B; meanwhile, ChatGPT obtains a FactCC score of 0.857. Although ChatGPT performs well in semantic understanding, its professionalism and domain specificity are slightly inferior to those of the specially optimized models. Notably, our plant protection question-answering system achieves the highest FactCC score of 0.897, demonstrating a significant advantage in ensuring the generated responses align with the reference facts.

Overall, the comprehensive analysis demonstrates that integrating fine-tuning technology with an external knowledge base substantially enhances domain-specific performance. Our plant protection question-answering system excels in both accuracy and professionalism, validating the effectiveness of this approach. Moreover, LoRA shows significant potential in improving the domain adaptability of large language models, offering valuable insights for constructing domain-specific systems. In contrast, general-purpose models exhibit notable limitations in specialized tasks and require further fine-tuning or knowledge enhancement.

These experimental results confirm that meticulously fine-tuning a large language model on a high-quality plant protection dataset—combined with an external professional knowledge base—can markedly optimize performance, while also providing a novel technical perspective and development direction for domain-specific question-answering systems.

4.3. Ablation Experiment

We further validate the effectiveness of fine-tuning large language models and integrating local knowledge bases through ablation experiments. Fine-tuning aims to adapt large language models to achieve enhanced performance in specific domains, enabling the model to capture domain-specific information more accurately and improve the quality of its responses. Consequently, evaluating the performance of the fine-tuned model is essential. Similarly, integrating a local knowledge base enhances answer accuracy and professionalism by incorporating domain-specific resources, which, in turn, improves system stability and credibility—thus necessitating performance verification.

In our ablation experiments, we employed a plant protection knowledge evaluation dataset to assess the impact of these two functions. Specifically, we removed the fine-tuning and knowledge base modules individually and measured the resulting changes in ROUGE, BERTScore, and FactCC.

As shown in Table 3, progressively removing the fine-tuning and knowledge base modules significantly degrades the performance of the plant protection question-answering system. In the complete system, ROUGE-1, ROUGE-2, and ROUGE-L scores are 0.578, 0.321, and 0.377, respectively, and the BERTScore F1 Score is 0.970, indicating strong performance in both generation quality and semantic consistency.

Table 3. Ablation analysis of LoRA and RAG on the performance of the plant protection QA system.

Models	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**BERTScore**	**FactCC**
				F1 Score	**F1 Score**
InternLM-20B	0.324	0.121	0.200	0.936	0.829
InternLM-20B + RAG	0.487	0.273	0.304	0.960	0.859
InternLM-20B + LoRA	0.394	0.139	0.254	0.953	0.846
Plant Protection QA System	**0.578**	**0.321**	**0.377**	**0.970**	**0.897**

Note: the best results are highlighted in bold.

When the fine-tuning module is removed, ROUGE-1, ROUGE-2, and ROUGE-L decline to 0.487, 0.273, and 0.304, respectively, while the BERTScore metrics drop to 0.960. These changes demonstrate that fine-tuning significantly enhances the model's keyword coverage, language expression accuracy, and semantic relevance. Upon further removal of the knowledge base module, ROUGE-1 falls to 0.394, ROUGE-2 to 0.139, and ROUGE-L to 0.254, while the BERTScore F1 Score decreases to 0.953. The experimental results show that the FactCC F1 scores varied significantly when different techniques were applied to the baseline model. The baseline model, InternLM-20B, achieved a FactCC F1 score of 0.829. By implementing LoRA alone, this metric improved to 0.846, confirming that effective parameter tuning can enhance factual consistency and domain-specific performance. Notably, introducing the external knowledge base module alone to the baseline model increased the score to 0.859. This highlights that while fine-tuning remains crucial, the local knowledge base also plays a vital role in ensuring factual accuracy. This result underscores the substantial contribution of the local knowledge base to improving the professionalism and accuracy of system responses, particularly by providing more comprehensive and credible content.

Compared to the pre-trained InternLM-20B, the plant protection question-answering system exhibits substantially improved performance, validating the necessity and effectiveness of both fine-tuning and local knowledge base integration. The combination of these enhancements not only boosts model performance in specific domains but also highlights the value of a modular design in domain-specific question-answering systems.

5. Discussion

The research on question-answering systems in the field of plant protection has long attracted significant attention. Traditional agricultural information services have predominantly relied on manual expert consultation and the use of scattered literature resources. As a result, when farmers are confronted with challenges such as pest control, disease management, and crop cultivation, it is often difficult for them to obtain accurate and comprehensive information promptly. This issue becomes even more critical in the context of modern agriculture, where rapid changes in planting structures and production environments have led to an increasing demand for both specialized knowledge and real-time data. Consequently, the urgent task at hand is to harness the latest advancements in artificial intelligence to construct an intelligent question-answering system that is both efficient and accurate.

When researchers began applying general large-scale pre-trained language models to specialized domains, they quickly discovered several shortcomings. On one hand, because the corpora used during pre-training are predominantly sourced from general domains, these models do not acquire sufficient specialized knowledge in plant protection. This lack of domain-specific understanding often results in misinterpretations when processing agricultural queries. On the other hand, these models are prone to a phenomenon known as “hallucination” during answer generation, whereby the content produced may not correspond to facts, thereby misleading users. Furthermore, agricultural information is inherently dynamic and exhibits pronounced regional differences. When dealing with diverse geographical environments and varying crop types, general-purpose models frequently fall short in providing personalized suggestions that are tailored to local conditions. The issue of data scarcity further exacerbates the problem; the limited availability of accurate plant protection question–answer pairs makes it challenging to fully capture the intricate domain characteristics during the model fine-tuning process, ultimately affecting overall performance and practical usability.

In response to these challenges, this study proposes a method that combines large language model fine-tuning with RAG techniques. In terms of large language model fine-tuning, the study employs LoRA (low-rank adaptation). By performing a low-rank adaptation on a subset of the parameters within the pre-trained model, LoRA enables the efficient injection of specialized plant protection knowledge without incurring significant computational overhead. This parameter-efficient fine-tuning method is designed to retain the rich linguistic capabilities that the model has acquired from a general corpus while simultaneously enhancing its accuracy and professionalism in answering domain-specific questions.

Moreover, to address the issues of knowledge loss and the “hallucination” phenomenon that may occur during the answer generation process, the proposed system incorporates RAG technology. During the answer generation phase, the system is capable of retrieving information relevant to the query from professional literature and agricultural databases in real time. It then integrates this external knowledge with the model’s internal generation results. Such an approach not only mitigates the risk of generating erroneous content but also enhances the authority and credibility of the produced answers. To further support the fine-tuning process, this study has constructed a comprehensive plant protection dataset containing over 9000 question–answer pairs. This dataset encompasses both fundamental theoretical knowledge and rigorously validated professional content extracted from academic publications. It serves as a high-quality training resource that enables the system to deliver more precise responses when faced with a wide range of complex agricultural issues.

In the experimental phase, the performance of multiple models was compared on plant protection question-answer tasks. These included a general large language model that had not undergone any fine-tuning, models that were fine-tuned using LoRA, and systems that incorporated the RAG technique in addition to fine-tuning. Comprehensive evaluations were conducted using a combination of expert reviews and automated evaluation metrics. The experimental results demonstrated that the model optimized with both LoRA and RAG integration exhibited significant improvements in answer accuracy, domain professionalism, and overall interpretability. Specifically, when addressing questions related to agricultural pest control, crop growth regulation, and the promotion of regional agricultural technologies, the optimized system was able to better capture key information within the queries. By effectively integrating real-time retrieved external professional data, the system generated responses that were both accurate and practically informative. This not only validates the effectiveness of parameter-efficient fine-tuning in specialized applications but also highlights the positive impact of real-time information retrieval on reducing "hallucination" problem frequently encountered by large-scale models. In several instances, the performance of individual question–answer pairs was particularly noteworthy, suggesting that the proposed method has substantial potential to enhance system robustness and domain adaptability.

Despite these promising outcomes, there remain several shortcomings in the present research. First, although the constructed plant protection dataset covers a substantial amount of foundational and professional knowledge, its breadth and depth are still limited. Certain specialized issues in peripheral subdomains have not been fully captured, which may result in suboptimal model performance when such edge-case queries arise. Second, the effectiveness of the RAG technique is, to a considerable extent, contingent upon the quality and update frequency of the external literature databases it relies upon. If the external knowledge sources are not updated promptly or lack sufficient authority, the accuracy and reliability of the generated answers could be adversely affected. Third, the current research has primarily focused on processing textual data, without exploring the integration of multimodal information. This limitation means that the system's applicability is restricted when it comes to handling queries that involve multiple forms of data, such as images, videos, or audio. Given that multimodal question–answer systems may play an increasingly important role in future agricultural information services, the challenge of effectively incorporating multimodal data into such systems represents a crucial direction for future research. Finally, although this study has achieved notable advancements in the areas of large language model fine-tuning and retrieval-augmented generation, there remains significant room for further optimization in terms of injecting domain-specific knowledge and implementing robust real-time information update mechanisms. These aspects warrant continuous improvement and further exploration in subsequent research endeavors.

6. Conclusions

This paper presents the development of an intelligent question-answering system tailored for the plant protection domain, leveraging large language model fine-tuning and retrieval-augmented generation (RAG). Extensive experiments demonstrate the efficacy of this approach in enhancing answer accuracy, professionalism, and interpretability. By employing low-rank adaptation (LoRA), the pre-trained model is efficiently infused with specialized domain knowledge without a significant increase in computational overhead. Simultaneously, the integration of RAG, which facilitates real-time retrieval of professional literature and agricultural data, markedly reduces the occurrence of "hallucinated" information, thereby bolstering the credibility of the responses. Comparative analyses reveal that post fine-tuning and real-time retrieval optimization, the system exhibits robust performance and delivers high-quality answers to challenges such as agricultural pest control, crop growth regulation, and regional agricultural promotion.

Nonetheless, certain limitations persist within this research. The constructed plant protection dataset, while encompassing fundamental theories and specialized knowledge, requires further expansion in both breadth and depth, particularly concerning niche areas. The system's heavy reliance on external literature repositories means that the quality and update frequency of these sources directly influences response efficacy. Moreover, the current focus on textual data processing, without the integration of multimodal information like images and videos, may constrain the system's applicability in future agricultural contexts. Despite these challenges, this study offers a viable technical pathway for domain-specific question-answering systems, underscoring the significant role of combining large language model fine-tuning with real-time information retrieval in reducing "hallucination" issues and enhancing domain adaptability.

Future endeavors should aim to expand the dataset's scope, enhance the authority of external knowledge bases, and explore the fusion of multimodal data. Such efforts will serve to refine system functionality and elevate overall performance. In essence, this work introduces innovative perspectives and practical insights for constructing intelligent question-answering systems in the plant protection sector, contributing positively to the advancement of agricultural informatization and informed decision-making.

Author Contributions: Conceptualization, J.X.; methodology, J.X., L.P. and Y.L.; software, J.X., L.P. and Y.L.; validation, J.X., L.P. and Y.L.; formal analysis, L.Z. (Lei Zhu), L.Z. (Lizhuo Zhang) and S.T.; investigation, L.P.; resources, J.X., L.Z. (Lei Zhu), L.Z. (Lizhuo Zhang) and S.T.; data curation, Y.L.; writing—original draft, J.X.; writing—review and editing, J.X., L.P., Y.L., L.Z. (Lei Zhu), L.Z. (Lizhuo Zhang) and S.T.; visualization, J.X.; supervision, L.Z. (Lei Zhu), L.Z. (Lizhuo Zhang) and S.T.; project administration, S.T.; funding acquisition, L.Z. (Lei Zhu), L.Z. (Lizhuo Zhang) and S.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No. 62202163); the Research Foundation of Education Bureau of Hunan Province, China (Grant No. 22A0145); and the Research Foundation of Science and Technology Bureau of Hunan Province, China (Grant No. 2023NK2011).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study can be obtained upon request from the corresponding author. Other datasets can be obtained from their original papers.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NLP	Natural language processing
LLM	Large language model
LoRA	Low-rank adaptation
RAG	Retrieval-augmented generation
FAISS	Facebook AI Similarity Search
BERT	Bidirectional Encoder Representations from Transformers
GPT	Generative Pre-trained Transformer
GPT-3	Generative Pre-trained Transformer 3
TREC-8	Text Retrieval Conference

SMS	Short Message Service
FAQs	Frequently Asked Questions
BLEU	Bilingual Evaluation Understudy
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
BERTscore	Bidirectional Encoder Representations from Transformers score

References

1. Alston, J.M.; Pardey, P.G. Agriculture in the Global Economy. *J. Econ. Perspect.* **2014**, *28*, 121–146. [CrossRef]
2. Lucas, J.A. Advances in plant disease and pest management. *J. Agric. Sci.* **2011**, *149* (Suppl. S1), 91–114. [CrossRef]
3. Klauser, D. Challenges in monitoring and managing plant diseases in developing countries. *J. Plant Dis. Prot.* **2018**, *125*, 235–237. [CrossRef]
4. Chakraborty, S.; Newton, A.C. Climate change, plant diseases and food security: An overview. *Plant Pathol.* **2011**, *60*, 2–14. [CrossRef]
5. Hadi, M.U.; Qureshi, R.; Shah, A.; Muneer, A.; Irfan, M.; Zafar, A.; Shaikh, M.B.; Akhtar, N.; Wu, J.; Mirjalil, S.I. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Prepr.* **2023**, *3*, 1–29. [CrossRef]
6. Patil, R.; Gudivada, V. A review of current trends, techniques, and challenges in large language models (llms). *Appl. Sci.* **2024**, *14*, 2074. [CrossRef]
7. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event, Canada, 3–10 March 2021; pp. 610–623. [CrossRef]
8. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
9. Wortsman, M.; Ilharco, G.; Kim, J.W.; Li, M.; Kornblith, S.; Roelofs, R.; Lopes, R.G.; Hajishirzi, H.; Farhadi, A.; Namkoong, H.; et al. Robust fine-tuning of zero-shot models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 7959–7971. [CrossRef]
10. Kumar, A.; Raghunathan, A.; Jones, R.; Ma, T.; Liang, P. Fine-Tuning Can Distort Pretrained Features and Underperform Out-of-Distribution. In Proceedings of the 10th International Conference on Learning Representations, Virtual Event, 25–29 April 2022. [CrossRef]
11. dos Santos Junior, J.C.; Hu, R.; Song, R.; Bai, Y. Domain-Driven LLM Development: Insights into RAG and Fine-Tuning Practices. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Barcelona, Spain, 25–29 August 2024; pp. 6416–6417. [CrossRef]
12. Mishra, A.; Jain, S.K. A survey on question answering systems with classification. *J. King Saud Univ. Comput. Inf. Sci.* **2016**, *28*, 345–361. [CrossRef]
13. Green, B.F., Jr.; Wolf, A.K.; Chomsky, C.; Laughery, K. Baseball: An automatic question-answerer. In Proceedings of the Western Joint IRE-AIEE-ACM Computer Conference, Los Angeles, CA, USA, 9–11 May 1961; pp. 219–224. [CrossRef]
14. Woods, W.A. Lunar Rocks in Natural English: Explorations in Natural Language Question Answering. In *Linguistic Structures Processing*; Zampolli, A., Ed.; Linguistic Structures Processing: Amsterdam, The Netherlands, 1977; pp. 521–569.
15. Wilensky, R. The Berkeley UNIX consultant project. In *Wissensbasierte Systeme: 2. Internationaler GI-Kongreß München*; Springer: Berlin/Heidelberg, Germany, 1987; pp. 286–296. [CrossRef]
16. Ferrucci, D.; Brown, E.; Chu-Carroll, J.; Fan, J.; Gondek, D.; Kalyanpur, A.A.; Lally, A.; Murdock, J.W.; Nyberg, E.; Prager, J.; et al. Building Watson: An overview of the DeepQA project. *AI Mag.* **2010**, *31*, 59–79. [CrossRef]
17. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [CrossRef]
18. Floridi, L.; Chiriatti, M. GPT-3: Its nature, scope, limits, and consequences. *Minds Mach.* **2020**, *30*, 681–694. [CrossRef]
19. Jain, M.; Kumar, P.; Bhansali, I.; Liao, Q.V.; Truong, K.; Patel, S. FarmChat: A conversational agent to answer farmer queries. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2018**, *2*, 1–22. [CrossRef]
20. Tende, I.G.; Aburada, K.; Yamaba, H.; Katayama, T.; Okazaki, N. Proposal for a crop protection information system for rural farmers in tanzania. *Agronomy* **2021**, *11*, 2411. [CrossRef]
21. Koopman, B.; Mourad, A.; Li, H.; van der Vegt, A.; Zhuang, S.; Gibson, S.; Dang, Y.; Lawrence, D.; Zuccon, G. AgAsk: An agent to help answer farmer's questions from scientific documents. *Int. J. Digit. Libr.* **2024**, *25*, 569–584. [CrossRef]

22. Zhang, D.; Chen, X.; Zhang, Y.; Qin, S. Template-based Chatbot for agriculture related FAQs. *arXiv* **2021**, arXiv:2107.12595. [CrossRef]
23. Dhavale, C.; Pawar, T.; Singh, A.; Pole, S.; Sabat, K. Revolutionizing farming: Gan-enhanced imaging, cnn disease detection, and llm farmer assistant. In Proceedings of the 2024 2nd International Conference on Computer, Communication and Control (IC4), Indore, India, 8–10 February 2024; IEEE: New York, NY, USA, 2024; pp. 1–6.
24. Klair, Y.S.; Agrawal, K.; Kumar, A. Impact of generative ai in diagnosing diseases in agriculture. In Proceedings of the 2024 2nd International Conference on Disruptive Technologies (ICDT), Greater Noida, India, 5–16 March 2024; IEEE: New York, NY, USA, 2024; pp. 870–875.
25. Madaan, V.; Bindal, G.; Singh, S.; Yadav, S.K.; Singh, A.; Sinha, P.; Nagpal, D. Integrating language models and machine learning for crop disease detection for farmer guidance. In Proceedings of the Workshop on Advances in Computational Intelligence (ACI-2023) Co-Located with the 2nd International Conference on Artificial Intelligence and Data Science (ICAIDS-2023), Hyderabad, Telangana, India, 29–30 December 2023; pp. 29–30.
26. Majumder, S.; Khandelwal, Y. Computer vision and Generative AI for yield prediction in Digital Agriculture. In Proceedings of the 2024 2nd International Conference on Networking and Communications (ICNWC), Chennai, India, 2–4 April 2024; IEEE: New York, NY, USA, 2024; pp. 1–6.
27. Fahim-Ul-Islam, M.; Chakrabarty, A.; Ahmed, S.T.; Rahman, R.; Kwon, H.H.; Piran, M.J. A Comprehensive Approach Towards Wheat Leaf Disease Identification Leveraging Transformer Models and Federated Learning. *IEEE Access* **2024**, *12*, 109128–109156.
28. Li, W.; Yu, L.; Wu, M.; Liu, J.; Hao, M.; Li, Y. DoctorGPT: A Large Language Model with Chinese Medical Question-Answering Capabilities. In Proceedings of the 2023 International Conference on High Performance Big Data and Intelligent Systems (HDIS), Macau, China, 6–8 December 2023; IEEE: New York, NY, USA, 2023; pp. 186–193. [CrossRef]
29. Honglin, X.; Sheng, W.; Yitao, Z.; Zhao, Z.; Liu, Y.; Huang, L.; Wang, Q.; Shen, D. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv* **2023**, arXiv:2304.01097. [CrossRef]
30. Wenzek, G.; Lachaux, M.-A.; Conneau, A.; Chaudhary, V.; Guzmán, F.; Joulin, A.; Grave, E. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, 13–15 May 2020; European Language Resources Association: Paris, France, 2020; pp. 4003–4012.
31. Lee, K.; Ippolito, D.; Nystrom, A.; Zhang, C.; Eck, D.; Callison-Burch, C.; Carlini, N. Deduplicating Training Data Makes Language Models Better. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 22–27 May 2022; Association for Computational Linguistics: Dublin, Ireland, 2022; pp. 8424–8445. [CrossRef]
32. Alberti, C.; Andor, D.; Pitler, E.; Devlin, J.; Collins, M. Synthetic QA Corpora Generation with Roundtrip Consistency. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; Association for Computational Linguistics: Dublin, Ireland, 2019; pp. 6168–6173. [CrossRef]
33. Lee, S.; Kim, H.; Kang, J. LIQUID: A framework for list question answering dataset generation. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 13014–13024. [CrossRef]
34. InternLM Team. Internlm: A Multilingual Language Model with Progressively Enhanced Capabilities [EB/OL]. 27 September 2023.
35. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Chen, W. Lora: Low-rank adaptation of large language models. *ICLR* **2022**, *1*, 3. [CrossRef]
36. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-T.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474. [CrossRef]
37. Zhao, P.; Zhang, H.; Yu, Q.; Wang, Z.; Geng, Y.; Fu, F.; Yang, L.; Zhang, W.; Jiang, J.; Cui, B. Retrieval-augmented generation for ai-generated content: A survey. *arXiv* **2024**, arXiv:2402.19473. [CrossRef]
38. Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 1–45. [CrossRef]
39. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; Association for Computational Linguistics: Dublin, Ireland, 2002; pp. 311–318. [CrossRef]
40. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*; Association for Computational Linguistics: Barcelona, Spain, 2004; pp. 74–81.
41. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating Text Generation with BERT. In Proceedings of the International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26–30 April 2020. [CrossRef]
42. Wang, A.; Cho, K.; Lewis, M. Asking and answering questions to evaluate the factual consistency of summaries. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; Association for Computational Linguistics: Dublin, Ireland, 2020; pp. 5008–5020.

43. Kryściński, W.; McCann, B.; Xiong, C.; Socher, R. Evaluating the factual consistency of abstractive text summarization. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; Association for Computational Linguistics: Dublin, Ireland, 2020; pp. 9332–9346.
44. Durmus, E.; He, H.; Diab, M. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; Association for Computational Linguistics: Dublin, Ireland, 2020; pp. 5055–5070.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.