Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# LoRA-Based Instruction Finetuning for Agronomic Knowledge Graph Triple Extraction and GraphRAG-Based Retrieval

**Rohini Kokare**

rbk22.comp@coeptech.ac.in

COEP Technological University

**Sunil Mane**

COEP Technological University

**Research Article**

**Keywords:** Instructional Fine-Tuning, Low-Rank Adaptation, triplet extraction, GraphRAG, Knowledge Graph, Large Language Models

**Posted Date:** June 1st, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-9601800/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

**Version of Record:** A version of this preprint was published at Discover Artificial Intelligence on October 7th, 2026. See the published version at https://doi.org/10.1007/s44163-026-02410-w.

# LoRA-Based Instruction Finetuning for Agronomic Knowledge Graph Triple Extraction and GraphRAG-Based Retrieval

Rohini Kokare[1*] and Sunil Mane[2]

[1*,2]Department of Computer Science and Engineering, COEP Technological University, Wellesley Rd, Shivajinagar, Pune, 411005, Maharashtra, India.

*Corresponding author(s). E-mail(s): rbk22.comp@coeptech.ac.in, sunilbmane.comp@coeptech.ac.in;

**Abstract**

Due to massive growth in the unstructured agricultural data generated from reports, research articles, news, farmers' feedback, experts' advice etc, there is a need to utilize this information efficiently for knowledge representation and decision making. There are multiple techniques where information extraction and managing it in a structured manner can be less challenging. LLMs are used widely for extracting meaningful information in the way the user instructs. More precisely, LLMs when fine-tuned on domain-specific data e.g, agricultural data helps to retrieve information promptly. We propose a framework where the Llama model is fine-tuned using LoRA technique and instruction dataset for extracting triplets from the unstructured agronomic data. Further this Llama-LoRA-based model is used to generate knowledge which will efficiently help to retrieve the query results to user using GraphRAG. Domain specific instructional fine-tuning has shown significant rise in the performance. ROUGE-Precision and BLEU-precision results of the proposed framework has outperformed the response retrieval compared to other retrieval techniques. Proposed results of domain-specific instructional fine-tuned Llama-LoRA model has overcome the limitations of knowledge representation and unstructured data extraction through GraphRAG resulting in semantic search with contextual retrieval.

**Keywords:** Instructional Fine-Tuning, Low-Rank Adaptation, triplet extraction, GraphRAG, Knowledge Graph, Large Language Models

# 1 Introduction

Now-a-days, huge amount of unstructured data has been generated and it is exponentially growing in all the application domains. The data that are continuously generated (research data, experimental results, real time results, historical evidence, etc.) in different formats, modalities, and templates etc., from domains like agronomics should be used for further analysis and evaluations. However, most of this data remains locked in unstructured text formats. While using this information is crucial to the advancement of precision agriculture and automated decision-support systems, converting fragmented text into structured, actionable knowledge poses a significant computational challenge.

Traditionally, Domain specific information extraction (IE) requires a complex pipeline which comprises a specific step-by-step process flow. These conventional methods separately depend on Named Entity Recognition(NER) and Relation Extraction (RE). They are not only labor intensive to train, but are also susceptible to cascading errors. Recently, Large Language Models (LLMs) have shown remarkable abilities in text extraction and response generation through domain specific prompts. Despite LLMs being powerful in the field of NLP, they often struggle to accurately map complex, multi-way relationships. Moreover, in the context of agriculture, We have a huge amount of unstructured information available like crop protection, crop disease, crop varieties, farm mechanization, crop production, etc.. Extracting semantic knowledge, fact finding and LLMs hallucination poses a significant risk in information extraction.

Nevertheless, out-of-the-box LLMs often lack the domain-specific precision needed for structured information extraction tasks such as generating subject–predicate–object (SPO) triplets. To overcome this limitation, instructional fine-tuning has evolved as an effective approach where pre-trained models adapt to specialized domains using task-specific prompts and curated datasets.

Traditional VectorDB retrieval and RAG-based retrieval often fail while extracting contextual and semantic knowledge. Extracting inherent relations and facts while answering the complex domain specific queries, where response relies on the multi-hop response generation, takes significant efforts. For this reason, when LLM responses are blended with GraphRAG retrieval, provides a better alternative to conventional retrieval mechanisms. Implementing GraphRAG is an automated, high-fidelity extraction technique where the raw text is transformed into a triplet of subject-predicate-object.

In this paper, We propose an integrated full length structure where GraphRAG helps to precisely produce an outcome, that links unstructured agronomic text and contextual retrieval system. Generalized prompting or full parameter LLM fine-tuning is computationally inefficient. Depending on such techniques is time consuming. So, PEFT is a supervised LLM fine-tuning technique where it focuses only on some required parameters, while all other parameters are frozen. We acquire a pre-trained Llama model and employ Low Rank Adaptation (LoRA) with instructional fine-tuning on the specific task of triple extraction.

When the proposed framework is optimized structurally, on a curated dataset, the tuned model will learn precisely for the agronomic domain. The triplets generated from LLM are then pushed onto the Neo4j graph database, where the knowledge

graph is constructed using multi-relational information. Ultimately, we illustrate how this tailored KG acts as the core retrieval mechanism for a GraphRAG system, facilitating highly precise, context-sensitive, and hallucination-free queries of agricultural information.

This study contributes to the following significant outcomes:

1. Targeted instructional fine-tuning: Specific modules of the pre-trained LLM architecture, including attention and MLP projection layers, are selected for adaptation and fine-tuning. The Llama model serves as the basis for this fine-tuning process. This will enhance the model for task specific and domain specific extraction of information. Instructional fine-tuning is performed using a domain-specific dataset comprising queries, triplets, and instructions.
2. End-to-end framework design: We outline a thorough, scalable framework, where the unstructured text data is processed and smoothly converted into a knowledge graph in Neo4j.
3. Domain-specific GraphRAG-based retrieval: We showcase the practical application of knowledge graph construction by integrating it into a GraphRAG system significantly boosting the accuracy and dependability of automated agricultural information retrieval. This allows a context-aware retrieval and enhanced response generation for user inquiries.

Further, the article is organized as follows: Section 2 analyzes the literature covered under the domain of KG retrieval using pre-trained LLMs and fine-tuned LLMs. Section 3 describes the LLM finetuning process and retrieval framework. Section 4 covers the experimental setup for fine-tuning. Our fine-tuned results and their analysis with other techniques are studied in Section 5. Section 6 concludes the study presented in this paper.

# 2 Background

## 2.1 Pre-trained LLMs

There are multiple methods for extracting triplets from the unstructured data. BERT-BiLSTM with CRF when integrated with the transformer model [1] to build a knowledge graph for journal related question answer tasks produces an efficient response for user questions. This technique improves the quality of responses and hence reliability of the system. During journal recommendation, in comparison with different combinations of models, fusion of BERT+BiLSTM+CRF improved performance marginally when evaluated for precision and accuracy. In educational domain, construction of knowledge graphs on Chinese contextual dataset [2] by just designing a structured prompt framework helps to extract the information. The prompt helps LLMs to follow the guidelines and accordingly extract information from the given dataset precisely. When multiple LLMs were experimented in comparison with the prompt framework, it showed an effective response generation.

In the field of communication systems, prompt engineering techniques can help fine-tune the model for task-specific systems. The entity and relationship extraction by employing DeepSeek model [3] and few-short prompting significantly improved

the performance as compared to other prompting combinations. Also to identify user behavioral patterns for a particular product depending on the product reviews and catalog by using LLM to generate a knowledge graph [4] can help in interaction of user and products.

### 2.2 Finetuned LLMs

LoRA-PEFT for the medical domain was proposed and a small language model was pre-trained. in the study in [5] extracted the medical text but it did not focus on structured data. The evaluation was totally computationally efficient like the response generation time, the tokens generated. Task specific fine-tuning was not explored. In the academic domain, [6] task specific LLM Finetuning for extraction of specific objectives and goals from the long thesis documents was studied using DoRA technique. The weights are decomposed into magnitude and direction. They are separately fine-tuned further using LoRA which is a base method. The evaluation showed this decomposition performed better in terms of semantic relevance from huge documents, hence generalized the extraction of unstructured domain specific text.

In QEFT, fine-tuning technique, the computational efficiency was considered for LLM during training. Authors in [7] experimented the gain in speed of training and demonstrated some specific attributes during fine tuning. The study notified the need for semantic understanding and extraction of data in structured form. The fusion of multiple LoRA adapters was used in [8] for LLM fine-tuning in the fashion domain. The combination of domain specific instructional fine-tuned FashionGPT-70B with multiple LoRA outperformed the SOTA techniques.

Instruction generation and task specific fine-tuning of Qwen [9] model was employed by using instruction dataset. The fine-tuned Qwen-IG model demonstrated effectiveness in task generation instructions in terms of higher complexity, accuracy in response generation, and more complex and diverse vocabulary. To precisely improve the models' response accuracy, instruction fine-tuning has shown a significant enhancement in its efficiency. In the field of mobile networking, fine-tuning [10] Llama on 5G real world network dataset has shown proficiency in the code generation of network analysis. The model out performed the GPT model with a good score of code generation.

## 3 Proposed Framework

Figure 1 illustrates the architecture of the proposed framework, detailing the process of LLM fine-tuning and retrieval utilizing knowledge graph retrieval-augmented generation (KG-RAG). Initially, the dataset is curated instructionally for triplet extraction which is trained using LoRA-finetuning technique for pre-trained Llama LLM model. This fine-tuned model is then tested for the inference. The triplets are extracted and pushed to Neo4j graph database for structured storage of linked nodes and entities. Further, GraphRAG is used to generate answers for user queries. This process is outlined in six stages:

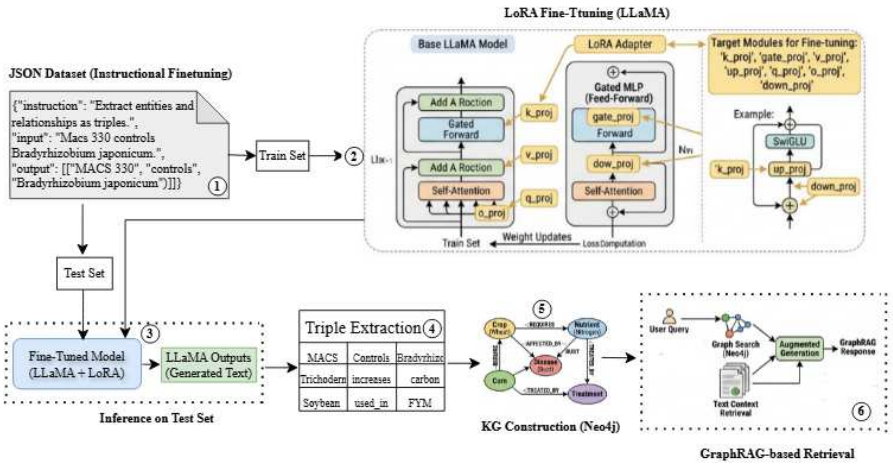



**Fig. 1** LLM-LoRA finetuning for information extraction and GraphRAG retrieval

## 3.1 Instructional Dataset Formation

In order to make LLM learn and adapt to domain-specific data for triplet extraction, the unstructured soybean agronomic data is converted into structured json format. The format is as follows: "Instruction" : "Extract entities and relationships as triples", "Input" : "Macs 330 controls Bradyrhizobium japonicum", "Output" : ["MACS 330", "controls", "Bradyrhizobium japonicum]"

Let the raw unstructured text corpus be denoted as $\mathcal{C}$. We formulate an instructional dataset as:

$$\mathcal{D} = \{(I, X_i, Y_i)\}_{i=1}^{N} \tag{1}$$

where:

- $I$ represents the static system instruction (e.g., "Extract entities and relationships as triples").
- $X_i \in \mathcal{C}$ is the input sequence of unstructured text.
- $Y_i$ is the target output, formally defined as a set of relational triples:

$$Y_i = \{(s_j, r_j, o_j)\}_{j=1}^{m} \tag{2}$$

where $s_j$ is the entity or subject, $r_j$ is the relation that connects the entities, and $o_j$ is the object.

The concatenated input sequence provided to the model is denoted as:

$$z_i = [I \oplus X_i] \tag{3}$$

where $\oplus$ denotes string concatenation.

### 3.2 Llama as a base model:

We integrated a pre-trained Llama model to fine-tune further on the attention(multiheaded, self) and feed forward layers of the network. The contextual representations are computed by the attention mechanism. This is estimated by considering the set of Query(Q), Key(K), Value(V) matrices and further scalar product is applied. The normalization of Query and Key matrix similarity is performed using the softmax function, and Values V are multiplied to get context aware responses. The weight metrics for each of Q,K,V ($W_q$, $W_k$, $W_v$) are used for projecting input X and further converted into $W_0$. After attention operation, a gated mechanism is used in the feed forward layer where the input is projected through $W_{gate}$, $W_{up}$, $W_{down}$ weight matrices through a non linear activation SiLU.. The activation of the attention layer and the formulation of all associated matrices are defined as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

Where, projections are defined as in equation 5.

$$Q = XW_q, \quad K = XW_k, \quad V = XW_v \tag{5}$$

$$O = \text{Attention}(Q, K, V)W_o \tag{6}$$

Feed-forward network (FFN) is projected as follows:

$$\text{FFN}(x) = \sigma(xW_{gate}) \odot (xW_{up})W_{down} \tag{7}$$

where:

- $W_q, W_k, W_v, W_o$: Attention projection matrices
- $W_{gate}, W_{up}, W_{down}$: FFN projection matrices
- $\sigma$: activation function (e.g., SiLU)

### 3.3 PEFT:

Low Rank Adaptation is a fine-tuning method that uses parameters efficiently by breaking down the adapter metrics into low-rank (r) matrices. The value of r can be as small as possible depending on the application (can be 2,4,8,16). If the size of decomposed matrix is n*r, where n=1 and r(rank)=1, the decomposed matrix calculation is as shown in fig. 2, when LoRA is used as PEFT. Instead of original weight matrix of size 512*512. LoRA decomposes into 512*1 and 1*512 which takes only 32,768 parameters, hence it is computationally efficient.

Our proposed model, used all target modules of attention and feed-forward layers.

$$\mathcal{T} = \{k_proj, q_proj, v_proj, o_proj, gate_proj, up_proj, down_proj\}$$

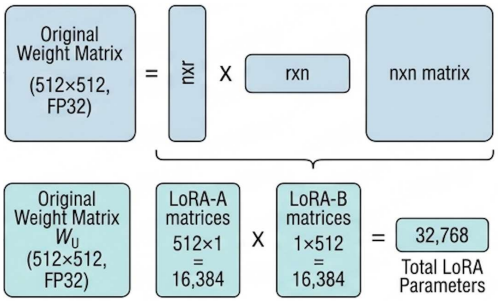



**Fig. 2** LoRA Decomposition

For a pre-trained frozen weight matrix $W \in \mathbb{R}^{d\times k}$, the low-rank adaptation matrix delta W, is decomposed into smaller matrices $A$ and $B$.

$$W' = W + \Delta W \tag{8}$$

$$\Delta W = BA \tag{9}$$

where,

- $B \in \mathbb{R}^{d\times r}, \quad A \in \mathbb{R}^{r\times k}$
- $r \ll \min(d, k)$ (low-rank constraint)

Hence, projected forward pass updates to:

$$h = x(W + BA) \tag{10}$$

Only $A$ and $B$ can be trained rest all are frozen, significantly reducing computational cost.

The main goal of the training process is to reduce the conditional likelihood i.e, probability of output that is correct and precise, when instructions and inputs are provided.

$$\mathcal{L}(\theta) = -\sum_{(I,X,Y)\in\mathcal{D}_{train}} \log P_\theta(Y \mid I, X) \tag{11}$$

## 3.4 Inference and Triplet Extraction

Structured output is generated after LoRA fine-tuning and an inference is checked on the test dataset. The structured output is in the form of triplets (subject, relation, object).

$$\hat{Y} = f_\theta(I, X) \tag{12}$$

A parsing function extracts triplets:

$$\hat{\mathcal{T}} = \text{Parse}(\hat{Y}) \tag{13}$$

Each triplet is represented as:

$$(s, r, o) \in \hat{\mathcal{T}} \tag{14}$$

### 3.5 KG Construction

The triplets are further pushed on to graph database and mapped to a structured graph representation as follows:

$$G = (V, E) \tag{15}$$

where:

- $V$ denotes the entity set, also referred to as nodes.
- $E$ represents the relation set, commonly known as edges.

Each triplet is transformed as:

$$(s, r, o) \rightarrow \left(v_s \xrightarrow{r} v_o\right) \tag{16}$$

Cypher query language is used to perform all the GraphDB operations like insert, delete, fetch certain nodes and retrieve information efficiently. We employed our previous method [11] to construct a knowledge graph from a fine-tuned model.

### 3.6 GraphRAG Retrieval

To enhance the query retrieval mechanism, the knowledge graph is then extended by using GraphRAG retrieval. The retrieval is more efficient using GraphRAG as it is the combination of LLM response generation and graph retrieval. The response is finally generated as follows:

$$G_q = \text{Retrieve}(G, q) \tag{17}$$

The final response is generated as:

$$y = f_\theta(q, G_q) \tag{18}$$

## 4 Experimental Setup

We considered the Indian agricultural use case of soybean crop. For the proposed problem statement, the dataset generated is scrapped from the websites like ICAR-National Soyabean Research Institute, Indore,[22] TNAU agritech portal [23]. These are unstructured documents with information like, crop protection, crop production, farm mechanization, good agricultural practices, variety information, region-wise recommended crops. We further processed and organized this dataset to help fine-tune domain-specific query retrieval systems with large language models..

We adopted the pre-trained Llama-3.2-1B-Instruct [24] model as the standard version for further calibration. This model is trained on 1 billion parameters using reinforcement learning and SFT (Supervised Fine Tuning) technique which has features like human feedback so as to get a hint of the generated response. It has multilingual

support for 8 languages. The pre-trained model is subsequently fine-tuned to address the requirements of a specific task. To analyze our Llama-LoRA fine-tuned model, the evaluation was performed based on the hyper-parameters given in table 1.

**Table 1** Llama-LoRA fine-tuned hyperparameters

| Hyperparameter | Value for finetuning |
|---|---|
| `per_device_train_batch_size` | 2 |
| `gradient_accumulation_steps` | 4 |
| `num_train_epochs` | 3 |
| `Learning_rate` | 2e-4 |
| `Max_length` | 512 |
| `Lora_rank` | 8 |
| `Lora_alpha` | 16 |
| `Lora_dropout` | 0.05 |
| `Target_modules` | ['k_proj', 'gate_proj', 'v_proj', 'up_proj', 'q_proj', 'o_proj', 'down_proj'] |

# 5 Results

We conducted an ablation study and evaluated our fine-tuned model with other techniques.

## 5.1 Impact of Instruction fine-tuning:

The importance of instruction-based prompts for fine-tuning is evaluated. As shown in Table 2, the prompt generates precise triplets according to the given instructions, rather than solely training the LLM with domain-specific data. Training the model with instructions enables adaptation to domain-specific triplet extraction. Instruction fine-tuning produces more accurate and specific facts, as well as consistent triplets, through the proposed solution.

We experimented target projection modules during LLM training using Attention gates (Q, K, V) and applied all layered gates with customized ranks for task-specific fine-tuning, using a one-shot prompt for entity and relation extraction. The results presented in Table 2 (Fig. 3) indicate that the module trained with all gates and instruction fine-tuning outperformed both attention-layer training and feed-forward-layer training.

## 5.2 End to end pipeline - Response generation:

We compared the generated responses to the ground truth and calculated ROUGE precision. The experiments demonstrated that the fine-tuned model, trained on a domain-specific triplet dataset, performed better than other response generation methods. Traditional vector-based retrieval, RAG-based retrieval, and knowledge graph-based retrieval were less efficient than the fine-tuned GraphRAG-based retrieval.

**Table 2** Analysis of Instructional Finetuning

| | Settings | LoRA Rank | Precision | Recall | F1-Score |
|---|---|---|---|---|---|
| 1 | Without Instruction-Attention Gates | 8 | 0.52 | 0.44 | 0.45 |
| 2 | Fusion (Customized Rank without instruction +Task specific prompt) | Attention=8, FFN=32 | 0.61 | 0.64 | 0.60 |
| 3 | Without Instruction-All Gates | 8 | 0.65 | 0.68 | 0.64 |
| 4 | With Instruction (Proposed) | 8 | **0.82** | **0.73** | **0.75** |

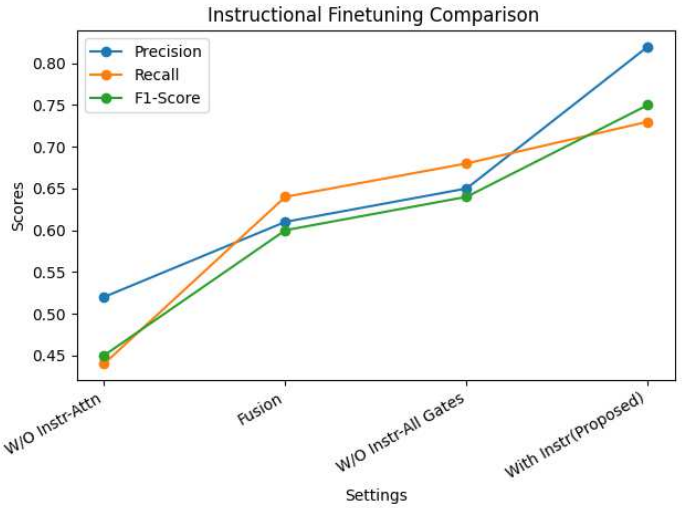



**Fig. 3** Instructional Finetuning Analysis

1. ROUGE Metric: This metric includes several variants, such as ROUGE-1, ROUGE-2, and ROUGE-L. This performance metric assesses the quality of response generated by the model. ROUGE-1 precision calculates the proportion of unigrams produced by the model that also appear in the reference response, relative to the total number of unigrams generated. ROUGE-2 precision evaluates the proportion of bigrams generated by the model that are present in the reference response, relative to the total number of bigrams generated. The ROUGE-L precision measures the length of the longest common subsequence between the model and reference responses, divided by the total number of unigrams generated. Table 3 demonstrates that the fine-tuned model achieves superior performance compared to other techniques. Figure 4 presents the scores of the fine-tuned model.
2. BLEU: BLEU precision represents a composite score that compares n-grams, ranging from unigrams to four-word sequences, against reference translations. This metric quantifies the degree to which the response generated by the model aligns with the reference answer. Table 4 presents the overall BLEU score as well as the four variants of the BLEU precision based on the length of n-grams. The results indicate that the fine-tuned model consistently achieves higher precision values than

**Table 3** ROUGE Precision Analysis across different techniques

| Technique | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| **VecDB** | 0.56 | 0.46 | 0.47 |
| **RAG** | 0.63 | 0.53 | 0.60 |
| **KGRAG** | 0.81 | 0.69 | 0.77 |
| **IFTLlama (Proposed)** | **0.90** | **0.74** | **0.84** |

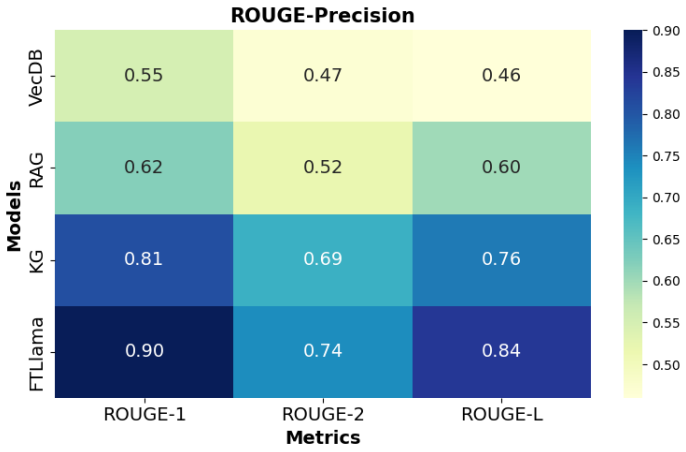



**Fig. 4** ROUGE Precision Analysis

all other retrieval models. Figure 5, a radar graph, clearly highlights that fine-tuned models perform better than the other techniques.

**Table 4** BLEU Precision Result Analysis across different techniques

| Technique | BLEU | Uni-gram | Bi-gram | Tri-gram | 4-gram | Avg N-gram |
|---|---|---|---|---|---|---|
| **VecDB** | 0.25 | 0.37 | 0.26 | 0.22 | 0.19 | 0.26 |
| **RAG** | 0.35 | 0.50 | 0.37 | 0.31 | 0.26 | 0.36 |
| **KGRAG** | 0.40 | 0.56 | 0.45 | 0.40 | 0.35 | 0.44 |
| **IFTLlama(Proposed)** | **0.52** | **0.88** | **0.76** | **0.71** | **0.67** | **0.75** |

# 6 Conclusion

Task-specific fine-tuning of LLM for agronomic applications enables the generation of focused responses to user queries. Fine-tuning mitigates the risk of overfitting by selectively updating matrix weights. We demonstrated that the Low-Rank Adaptation (LoRA) technique reduces computational requirements by training only specific

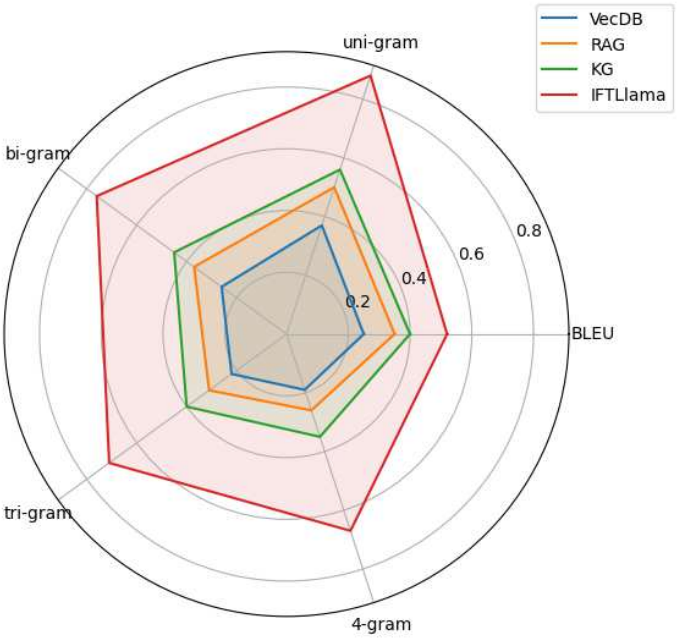



**Fig. 5** BLEU Precision Analysis

parameters of the pre-trained Llama model while freezing the remaining parameters. For task-specific training, we utilized a curated agronomic dataset comprising instructions, queries, and triplets. When applied for inference, the trained model effectively extracts triples and generates an agronomic knowledge graph, achieving superior retrieval precision compared to alternative methods by leveraging the GraphRAG technique for semantic and contextual retrieval. Our approach outperformed other state-of-the-art (SOTA) techniques. The fine-tuned Llama model for knowledge graph generation enables users to pose domain-specific queries in natural language, with responses generated by extracting nodes, relationships, and facts from the graph using GraphRAG. This efficient retrieval technique is expected to significantly benefit domain-specific users in question answering tasks.

In addition, we intend to extend this research to address more complex and multimodal unstructured elements. These elements will be transformed into structured knowledge graph stores utilizing multi-modal large language model (LLM) capabilities. The quality of the generated triplets following finetuning will be evaluated using specific metric measures.

## Acknowledgment

I would like to express my sincere gratitude to COEP Technological University, Pune, for providing the opportunity, resources, and support necessary to carry out this research. I am deeply thankful to my research guide, Dr Sunil. B Mane, for his

invaluable guidance and constant encouragement throughout this study [11].

**Data availability**: The dataset used in this study was created by the first corresponding author as part of the research work. Since it is not currently hosted in a public repository, interested researchers may request access from the first corresponding author.
**Conflict of interest**: There is no conflict of interest
**Competing interests:** The authors have no Competing interest to declare. All co-authors have seen and agree with the content of the manuscript and there is no financial interest to report.
**Funding Information**: No funding received for this research
**Author contribution**: This research was made possible by the collaboration and contributions of all authors
**Research Involving Human and/or Animals**: Not Applicable
**Informed Consent**: Not Applicable.
**Clinical Trial Registration**: This study is not a clinical trial. Hence, clinical trial registration details are not applicable.
**Consent to Publish**: Not applicable.
**Consent to Participate**: Not applicable.
**Ethics Approval**: Not applicable.

# References

[1] Zuo, Jiankun, and Jiaojiao Niu., "Construction of Journal Knowledge Graph Based on Deep Learning and LLM" Electronics 14, no. 9: 1728. 2025 https://doi.org/10.3390/electronics14091728

[2] J. Sun, Z. Zhang and X. He, "LLM4EduKG: LLM for Automatic Construction of Educational Knowledge Graph," 2024 International Conference on Networking and Network Applications (NaNA), Yinchuan City, China, 2024, pp. 269-275, doi: 10.1109/NaNA63151.2024.00051.

[3] Guo Chang, Jiaqi Liu, Wei Gao, Zhenhai Lu, Yao Li, Chengyuan Wang, and Jungang Yang, "A Large Language Model Driven Knowledge Graph Construction Scheme for Semantic Communication", Applied Sciences 15, no. 8: 4575. 2025 https://doi.org/10.3390/app15084575

[4] J. Vizcarra, S. Haruta and M. Kurokawa, "Representing the Interaction between Users and Products via LLM-assisted Knowledge Graph Construction," 2024 IEEE 18th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 2024, pp. 231-232, doi: 10.1109/ICSC59802.2024.00043.

[5] A. K. G. Rao, A. Jaggi and S. Naidu, "MEDFIT-LLM: Medical Enhancements through Domain-Focused Fine Tuning of Small Language Models," 2025 2nd International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE),

Chennai, India, 2025, pp. 1-5, doi: 10.1109/RMKMATE64874.2025.11042816.

[6] V. Tereshchenko, V. Martsinkevich, N. Aminov and A. Dukhanov, "The use of LLM to identify goals and objectives from the text of degree thesis using fine-tuning and prompt-engineering methods," 2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA), Antalya, Turkiye, 2025, pp. 1-5, doi: 10.1109/ACDSA65407.2025.11165914.

[7] Changhun Lee1, Jungyu Jin, Younghyun Cho, Eunhyeok Park, "QEFT: Quantization for Efficient Fine-Tuning of LLMs", Findings of the Association for Computational Linguistics: EMNLP 2024, pages 13823–13837 November 12-16, 2024 doi: 10.18653/v1/2024.findings-emnlp.811

[8] Dehong Gao, Yufei Ma, Sen Liu, Mengfei Song, Linbo Jin, Wen Jiang, Xin Wang, Wei Ning, Shanqing Yu, Qi Xuan, Xiaoyan Cai, Libin Yang, "FashionGPT: LLM instruction fine-tuning with multiple LoRA-adapter fusion", Knowledge-Based Systems, Volume 299, 2024, 112043, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2024.112043.

[9] Lu Zhang, Yu Liu, Yitian Luo, Feng Gao, and Jinguang Gu. "Qwen-IG: A Qwen-based Instruction Generation Model for LLM Fine-tuning", In Proceedings of the 2024 13th International Conference on Computing and Pattern Recognition (ICCPR '24), Association for Computing Machinery, New York, NY, USA, 2025, 295–302. https://doi.org/10.1145/3704323.3704357

[10] K. B. Kan, H. Mun, G. Cao and Y. Lee, "Mobile-LLaMA: Instruction Fine-Tuning Open-Source LLM for Network Analysis in 5G Networks," in IEEE Network, vol. 38, no. 5, pp. 76-83, Sept. 2024, https://doi.org/10.1109/MNET.2024.3421306

[11] R. Naik and S. Mane, "An Efficient Knowledge Graph Construction using LLM-Based Retrieval Augmented Generation for Agricultural Data," TENCON 2025 - 2025 IEEE Region 10 Conference (TENCON), Kota Kinabalu, Malaysia, 2025, pp. 155-159, doi: 10.1109/TENCON66050.2025.11375209

[12] S. S. Alahmari, L. O. Hall, P. R. Mouton and D. B. Goldgof, "Repeatability of Fine-Tuning Large Language Models Illustrated Using QLoRA," in IEEE Access, vol. 12, pp. 153221-153231, 2024, doi: 10.1109/ACCESS.2024.3470850.

[13] Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya, Poria1 Roy, Ka-Wei Lee, "LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models", Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 5254–5276 December 6-10, 2023 do: 10.18653/v1/2023.emnlp-main.319

[14] Veena G, Aparna Rajendran, Deepa Gupta, "Extracting Triplets from Domain-specific Texts using Fine-Tuned LLaMA Models: A Comparative Study", Procedia Computer Science, Volume 258, 2025, Pages 3750-3759, ISSN 1877-0509,

https://doi.org/10.1016/j.procs.2025.04.630.

[15] D.M. Anisuzzaman, Jeffrey G. Malins, Paul A. Friedman, Zachi I. Attia, "Fine-Tuning Large Language Models for Specialized Use Cases", Mayo Clinic Proceedings: Digital Health, Volume 3, Issue 1, 2025, 100184, ISSN 2949-7612, https://doi.org/10.1016/j.mcpdig.2024.11.005.

[16] Wu, X.-K.; Chen, M.; Li, W.; Wang, R.; Lu, L.; Liu, J.; Hwang, K.; Hao, Y.; Pan, Y.; Meng, Q.; et al. "LLM Fine-Tuning: Concepts, Opportunities, and Challenges", Big Data Cogn. Comput. 2025, 9, 87. https://doi.org/10.3390/bdcc9040087

[17] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, and Fei Wu. 2026. "Instruction Tuning for Large Language Models: A Survey", ACM Comput. Surv. 58, 7, Article 169 (May 2026), 36 pages. https://doi.org/10.1145/3777411

[18] S. Joshi, M. S. Khan, A. Dafe, K. Singh, V. Zope and T. Jhamtani, "Fine Tuning LLMs for Low Resource Languages", 2024 5th International Conference on Image Processing and Capsule Networks (ICIPCN), Dhulikhel, Nepal, 2024, pp. 511-519, doi: 10.1109/ICIPCN63822.2024.00090

[19] Q. Wang, C. Li, Y. Liu, Q. Zhu, J. Song and T. Shen, "An Adaptive Framework Embedded With LLM for Knowledge Graph Construction", in IEEE Transactions on Multimedia, vol. 27, pp. 2912-2923, 2025, doi: 10.1109/TMM.2025.3557717.

[20] Tuan Bui, Oanh Tran, Phuong Nguyen, Bao Ho, Long Nguyen, Thang Bui, and Tho Quan, "Cross-Data Knowledge Graph Construction for LLM-enabled Educational Question-Answering System: A Case Study at HCMUT", In Proceedings of the 1st ACM Workshop on AI-Powered Q&A Systems for Multimedia (AIQAM 24). Association for Computing Machinery, New York, 2024 NY, USA, 36–43. https://doi.org/10.1145/3643479.3662055

[21] Xinxin Liang, Zuoxu Wang, Mingrui Li, Zhijie Yan, "A survey of LLM-augmented knowledge graph construction and application in complex product design", Procedia CIRP, Volume 128, 2024, Pages 870-875, ISSN 2212-8271, https://doi.org/10.1016/j.procir.2024.07.069.

[22] https://icar-nsri.res.in/

[23] https://agritech.tnau.ac.in/

[24] https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct