F-LoRA-QA: Finetuning LLaMA Models with Low-Rank Adaptation for French Botanical Question Generation and Answering

Ayoub Nainia[1] **Régine Vignes-Lebbe**[1] **Hajar Mousannif**[2] **Jihad Zahir**[2,3]
[1]Institut de Systématique, Evolution, Biodiversité (ISYEB), Sorbonne Université, Muséum National d'Histoire Naturelle, CNRS, EPHE-PSL, Université des Antilles, F-75005, Paris, France
[2]LISI Laboratory, Cadi Ayyad University, Marrakesh, Morocco
[3]UMMISCO, IRD, France
ayoub.nainia@sorbonne-universite.fr,
regine.vignes_lebbe@sorbonne-universite.fr,
mousannif@uca.ac.ma, j.zahir@uca.ac.ma

Abstract

Despite recent advances in large language models (LLMs), most question-answering (QA) systems remain English-centric and poorly suited to domain-specific scientific texts. This linguistic and domain bias poses a major challenge in botany, where a substantial portion of knowledge is documented in French.

We introduce F-LoRA-QA, a fine-tuned LLaMA-based pipeline for French botanical QA, leveraging Low-Rank Adaptation (LoRA) for efficient domain adaptation. We construct a specialized dataset of 16,962 question-answer pairs extracted from scientific flora descriptions and fine-tune LLaMA models to retrieve structured knowledge from unstructured botanical texts.

Expert-based evaluation confirms the linguistic quality and domain relevance of the generated responses. Compared to baseline LLaMA models, F-LoRA-QA achieves a four-fold improvement in BLEU, a 70% ROUGE-1 F1 gain, a 16.8% increase in BERTScore F1, and an Exact Match improvement from 2.01% to 23.57%.

These results demonstrate the effectiveness of adapting LLMs to low-resource scientific domains and highlight the potential of our approach for automated trait extraction and biodiversity data structuring.

1 Introduction

As AI advances in biodiversity research, converting free-text morphological descriptions into structured, interoperable data remains a key challenge. In botany, detailed but unstructured identification keys hinder large-scale data curation, species comparison, and the integration of botanical knowledge into digital descriptor-based systems.

Recent advances in artificial intelligence, particularly in large language models (LLMs) and fine-tuning techniques, have significantly improved information extraction tasks such as Named Entity Recognition (NER), Relation Extraction (RE), and Question Answering (QA) (Touvron et al., 2023; Dagdelen et al., 2024). However, these advancements have primarily focused on general-domain texts and are overwhelmingly English-centric, leaving domain-specific and non-English applications underexplored. This linguistic and domain bias limits AI-driven breakthroughs in scientific disciplines such as botany, where a vast body of knowledge is documented in languages such as French.

French is a crucial language in botanical research, as many classical and contemporary flora documents, including taxonomic keys, species descriptions, and herbarium records, are written in French [1]. However, existing LLM-based QA systems are predominantly trained on English corpora, making them ineffective for processing such texts. Although prior efforts such as FloraNER (Nainia et al., 2024b) have focused on NER for species and morphological terms, they do not address question-answering or structured trait extraction. Consequently, there remains a critical gap in developing domain-adapted LLMs for French botanical QA.

Recent biodiversity NLP approaches have explored hybrid methods that combine rule-based heuristics with transformer-based models for relation extraction (Gabud et al., 2023). While these improve performance, they rely on hand-crafted rules and lack flexibility for open-ended QA. In contrast, LLMs have shown a strong potential for automating large-scale information extraction.

For instance, a recent study on ecological data extraction showed that LLMs can process scientific data over 50 times faster than human experts, achieving over 90% accuracy on categorical traits while struggling with complex quantitative values (Gougherty and Clipp, 2024). These findings highlight both the efficiency and limitations of LLMs

[1]https://inpn.mnhn.fr/informations/biodiversite/france

Proceedings of Recent Advances in Natural Language Processing,pages 787–796
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-091

in specialized domains, emphasizing the need for domain adaptation and structured validation.

Domain-specific benchmarks such as BioASQ (Krithara et al., 2023) and multilingual QA datasets like XQuAD (Artetxe et al., 2020) have been instrumental in evaluating scientific QA systems, but focus primarily on biomedical content and lack biodiversity-specific coverage.

Earlier trait extraction systems like FloraTraiter (Folk et al., 2024) relied on rule-based parsing and domain-specific heuristics, whereas recent LLM-based methods offer more flexible and scalable approaches (Marcos et al., 2025). However, these remain underexplored for biodiversity applications, particularly in low-resource languages such as French.

Despite growing accessibility, adapting LLMs to specialized fields like biodiversity and biomedicine is constrained by the high cost of full fine-tuning. Low-Rank Adaptation (LoRA) addresses this by introducing lightweight and trainable matrices, significantly reducing memory and computation costs while maintaining strong performance (Hu et al., 2021).

Previous research has explored domain adaptation for biomedical QA and NER, such as BioGPT (Luo et al., 2022) and AliBERT (Berhe et al., 2023), a French biomedical language model that outperforms general-purpose French LLMs like CamemBERT (Martin et al., 2020a) and FlauBERT (Le et al., 2020). However, these efforts are limited to biomedical tasks, often focus on classification or NER, and lack generalizable QA capabilities for biodiversity contexts.

To address these limitations, we propose **F-LoRA-QA**, a two-stage pipeline for French botanical QA, fine-tuned on a curated corpus of species descriptions. First, the system transforms a predefined list of standardized botanical traits (e.g., leaf shape, flower color) into natural language questions tailored to each species description. These questions are then used as prompts to extract the corresponding trait values from the free-text description. This approach enables the automated transformation of unstructured botanical text into structured character-state pairs.

A key application of this work lies in enriching descriptor-based systems such as Xper3 (Kerner et al., 2025), which rely on structured matrices to differentiate taxa. F-LoRA-QA aims to facilitate the semi-automated population of such descriptor models from legacy flora documents, supporting the development of interactive identification tools and enabling more efficient curation and integration of biodiversity data.

Here are the main contributions of our work:

1. We introduce the first fine-tuned LLaMA-based models for the generation and answering of botanical questions, adapted to French botanical texts.

2. We construct the first botanical Q&A instruction dataset for fine-tuning, comprising ≈17,000 (16,962) sample question-answer pairs with their botanical contexts, extracted from scientific flora documents.

3. We conduct a comprehensive expert-based evaluation, assessing our models on accuracy, completeness, fluency, and the use of botanical terminology to ensure both linguistic quality and domain relevance.

The remainder of this paper is structured as follows: Section 2 reviews related work on domain-specific LLM adaptation, Section 3 details our methodology for constructing the F-LoRA-QA pipeline, Section 4 presents our experimental setup and evaluation metrics, and Section 5 discusses the results, implications, and future directions.

2 Related Work

Recent efforts in biodiversity NLP have focused on extracting structured information from taxonomic literature using Named Entity Recognition (NER) and Relation Extraction (RE). TaxoNERD (Le Guillarme and Thuiller, 2022) applies deep learning to identify taxonomic entities in ecological texts, supporting species recognition in unstructured data. FloraNER (Nainia et al., 2024b) extends this to French botanical texts, introducing NER datasets for species and morphological terms and underscoring the need for multilingual adaptation.

While NER-based approaches like FloraNER and TaxoNERD enable species identification, they cannot provide structured answers to domain-specific questions. BiodivNERE (Abdelmageed et al., 2022) incorporates Relation Extraction (RE) to link entities, but is limited by predefined relation types. Hybrid approaches (Gabud et al., 2023) that combine rules and transformers improve RE accuracy, while recent work (Montero et al., 2024) uses LLaMA 2 and TaxoNERD for RE in English

and Spanish. However, these methods do not cover French taxonomic texts or explore question answering (QA) for more flexible information access.

In general, existing methods struggle with open-ended knowledge retrieval and require manual definition of relation and entity types. While NER and RE extract names and relations, they do not support structured comparison of species descriptions.

Departing from this, QA offers adaptive, context-aware responses that are particularly valuable for biodiversity NLP, where descriptions are rich but unstructured. Early domain-specific QA applications, such as ChatBBNJ (Wang et al., 2024), apply LLMs to biodiversity law. However, biodiversity-focused QA, especially for non-English taxonomic literature, remains largely unexplored.

Other works explore cost-efficient domain adaptation for QA. LeanContext (Arefeen et al., 2024) reduces inference costs by optimizing context selection in Retrieval-Augmented Generation (RAG), targeting input efficiency. This complements our use of LoRA (Low-Rank Adaptation), which focuses on parameter-efficient fine-tuning. While LeanContext reduces inference cost by selecting compact input contexts, LoRA minimizes training overhead by adapting only a subset of model parameters. Both strategies improve the efficiency of domain-specific QA, but at different stages of the pipeline. Similarly, FabricQA-Extractor (Wang and Fernandez, 2024) demonstrates the use of QA models for structured document extraction.

In contrast, traditional extractive QA models trained on datasets like SQuAD (Rajpurkar et al., 2016) rely on retrieval and lack generative flexibility. Our work leverages LLaMA fine-tuning to build a generative QA model capable of producing more context-aware and adaptive responses.

Adapting LLMs to domain-specific tasks has been a key focus in NLP, especially in biomedical and biodiversity contexts. BioGPT (Luo et al., 2022) showed that pre-training on specialized corpora improves performance for biomedical text generation and mining. Similarly, AliBERT (Berhe et al., 2023), a French biomedical model, outperforms general-purpose models like CamemBERT (Martin et al., 2020a) and FlauBERT (Le et al., 2020) on domain-specific tasks, highlighting the value of fine-tuning for non-English scientific texts.

Other biomedical LLMs, such as BioBERT (Lee et al., 2019), SciBERT (Beltagy et al., 2019), and PubMedBERT (Gu et al., 2021), further support the benefits of domain adaptation. However, similar efforts remain underexplored for biodiversity texts. Despite significant progress in biomedical NLP, particularly with ontology-linked entity linking using UMLS (Bodenreider, 2004) and MeSH[2], biodiversity NLP lacks a comparable infrastructure for taxonomic resources such as GBIF[3] and TAXREF[4], emphasizing the need for domain-specific LLM adaptation.

PLLaMa (Yang et al., 2024), an open-source LLaMA-2 model trained on over 1.5 million plant science documents, marks a major step toward generative plant knowledge retrieval. While effective on domain-specific QA tasks, it focuses primarily on English and does not address non-English taxonomic literature or morphological trait extraction.

Building on prior work, we show that LoRA-based fine-tuning can effectively adapt general-purpose LLMs for French botanical QA. To our knowledge, no LLM-based system currently retrieves French biodiversity knowledge. Our work addresses this gap through efficient adaptation of LLaMA for French botanical texts.

3 F-LoRA-QA Dataset Construction

The F-LoRA-QA dataset contains ≈17,000 samples (Table 1), where each sample consists of a **context**, a **question**, and an **answer**.

F-LoRA-QA Statistics	Value
Total Q&A Pairs for training	16,962
Total Unique Contexts	2,913
Avg. Q&A Pairs per Context	6
Total Q&A Pairs for evaluation	1697

Table 1: Statistics of the F-LoRA-QA dataset, including total contexts, questions, and answers.

The context serves as the foundational botanical knowledge, extracted from publicly available flora documents (Step 1 in Figure 1), including the FloraNER dataset (a dataset of morphological descriptions from Flora of New Caledonia) (Nainia et al., 2024a), *Flore de Madagascar et des Comores* (1974, 1976, 1994), *Flore du Cameroun* (1964, 1967, 1970, 1972, 1973, 1974), and *Flore du Gabon* (1962, 1968, 1983). These sources provide scientifically reliable, literature-based morpho-

[2] https://www.nlm.nih.gov/mesh/meshhome.html

[3] https://www.gbif.org/

[4] https://inpn.mnhn.fr/programme/referentiel-taxonomique-taxref

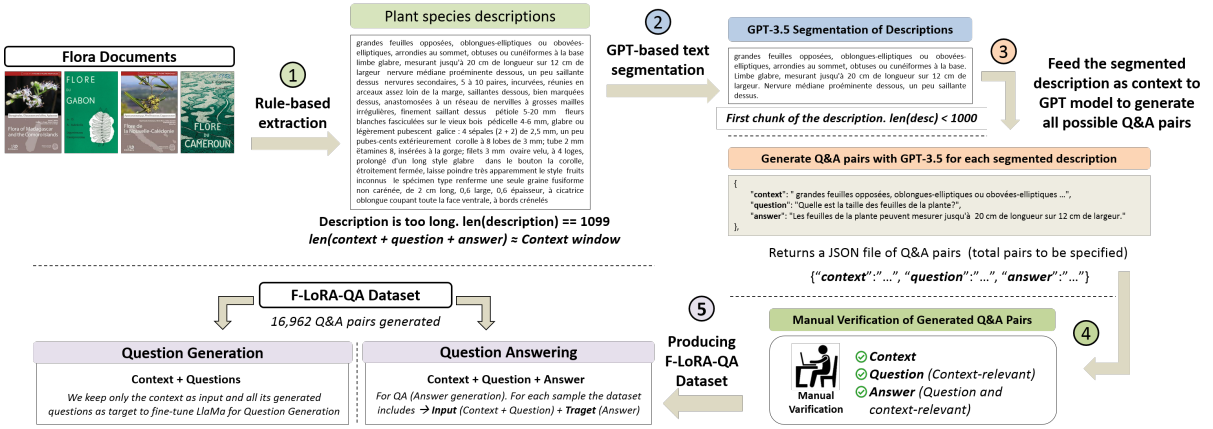

Figure 1: Process of segmenting botanical descriptions and generating Q&A pairs for the F-LoRA-QA dataset. The pipeline consists of four key steps: (1) Extracting plant species descriptions from flora documents, (2) Segmenting long descriptions into semantically coherent chunks using GPT-3.5, (3) Generating question-answer pairs from refined segments, and (4) Performing manual semantic verification to ensure that each generated question is answerable within its provided context. (5) The final dataset is structured for both question generation and question answering tasks.

logical descriptions of plant species, making them well-suited for training a botanical QA system.

To ensure robust evaluation, we curated a separate expert-based hold-out dataset of 1,697 examples, composed of botanical contexts exclusively extracted from *Flore du Sénégal*. This dataset was deliberately excluded from the fine-tuning corpus to evaluate the model's ability to generalize to unseen botanical texts from a distinct source.

Since botanical descriptions can be lengthy, directly using them as contexts could exceed the model's input limits, leading to excessive computational load. To address this, we segment the descriptions into semantically coherent spans of 300 to 1,000 characters using GPT-3.5 (Step 2 in Figure 1). These segments were purely extracted from the original descriptions and manually verified for coherence and botanical relevance. Segments of less than 300 characters were discarded to maintain content density and consistency (prompt provided in the supplementary material[5]).

Next, GPT-3.5 processed each segment independently to generate question-answer (QA) pairs (Step 3 in Figure 1). The model was prompted to generate all relevant QA pairs for the segment, with outputs structured in JSON format: each entry included the context, a list of questions, and their corresponding answers. The prompt (provided in the supplementary material[6]) encouraged diversity in the types of questions, including:

- **General feature questions** (e.g., *What are the key characteristics of this plant?*)
- **Specific detail questions** (e.g., *What is the shape and size of the leaves?*)
- **Structural relationship questions** (e.g., *What part of the plant is connected to the inflorescence?*)
- **Missing information questions** (e.g., *Are the fruits described?*)

The question-answer pairs form the unique entries in the dataset, which means that a single context can be associated with multiple question-answer pairs as shown in Table 2. Since botanical descriptions are highly detailed, they often contain multiple pieces of relevant information that allow for the generation of several questions per context. This structure ensures that the dataset captures various aspects of plant morphology, taxonomy, and botanical characteristics.

To ensure accuracy and domain relevance, the generated question-answer pairs undergo a manual verification process (Step 4 in Figure 1). Each question was reviewed to determine whether its subject was answerable using the given context and whether the corresponding answer correctly responded to the question. Question-answer pairs that did not meet these criteria were discarded. This verification step ensures that only contextually valid and botanically relevant pairs are retained in the dataset to make it reliable for fine-tuning.

[5] doi:10.5281/zenodo.16993669
[6] doi:10.5281/zenodo.16993669

The final F-LoRA-QA dataset was then restructured into two separate datasets (Step 5 in Figure 1): one for question generation, which retains only contexts and their corresponding questions (sample in Table 2), and another for question answering, where each sample consists of a context, a question, and its corresponding answer (sample in Table 3).

4 F-LoRA-QA Architecture

F-LoRA-QA is a two-stage pipeline for French botanical Question-Answering (QA), designed to extract standardized botanical traits from unstructured floristic descriptions. The system leverages Low-Rank Adaptation (LoRA) to fine-tune LLaMA 2-7B and LLaMA 3-8B models for domain-specific language understanding.

Our approach separates the task into two components: question generation and answer generation. The question generation model translates a predefined list of botanical traits (e.g., leaf shape, flower color) into natural language questions tailored to each species description. This serves both to improve trait retrieval and to support users who may lack domain expertise to formulate precise queries.

The answer generation model is trained to respond to these trait-specific questions using the corresponding species description as context. Importantly, it is fine-tuned on GPT-3.5-generated question-answer pairs that were verified by human experts to ensure consistency and high quality in the training data.

4.1 Stage 1: Question Generation

This stage focuses on generating a list of questions from botanical descriptions or any relevant botanical context provided. Given a botanical context x, the goal is to generate a structured sequence of questions $Q = (q_1, q_2, \ldots, q_n)$, where each question q_i is relevant to the information present in x (prompt provided in the supplementary material[7]).

To achieve this, we fine-tuned LLaMA 2-7B[8] and LLaMA 3-8B[9] using our F-LoRA-QA dataset, considering only context-question pairs. Each training sample consists of a botanical description (context) and its corresponding list of questions, as shown in Table 2.

The model is trained in a sequence-to-sequence framework (Sutskever et al., 2014), where it learns to generate multiple relevant questions for a given botanical passage. The model is optimized to output a structured list of questions in a single output sequence (Pan et al., 2020). At inference time, the question generation model can generate multiple questions per botanical passage, either as a batch or interactively in response to user queries.

Input - Botanical Description (Context)
Each pinna has (5-)7-13 pairs of opposite, sessile, leathery, subtrapezoidal or oblong, sometimes slightly arched leaflets. The blade is glabrous on both sides, with raised venation. The spike-like cluster inflorescences are densely tomentose and pedunculated, measuring 4 to 9 cm long.
Output - Questions
1. What is the length of the spike-like cluster inflorescences?
2. What is the leaflet blade like in terms of texture and venation?
3. Are the leaflets arched according to the text?
4. How many pairs of leaflets does each pinna have?

Table 2: Example of a training sample for the Question Generation stage. The model learns to generate a structured list of botanical questions from a given botanical description.

4.2 Stage 2: Answer Generation

In the second stage, the model is trained to generate answers based on both the botanical context and the question, as illustrated in Table 3. Here, the fine-tuned LLaMA model takes both the context and the generated question as input. The input sequence is structured as follows:

$$x = [\text{Context}] \oplus [\text{Question}] \quad (1)$$

where x represents the input text that is the concatenation of the botanical context with the question. The model then predicts a relevant answer in an autoregressive manner.

This task is formally framed as a conditional sequence generation problem, where the model learns the probability distribution over answer tokens given the input context and question. This follows the autoregressive decoding paradigm used in large-scale language models (Brown et al., 2020), where the model iteratively predicts tokens based

[7] doi:10.5281/zenodo.16993669
[8] LLaMA 2-7B, available on Hugging Face.
[9] LLaMA 3-8B, available on Hugging Face.

on input and previously generated sequence. Formally, our goal is to maximize the likelihood:

$$P(y \mid x) = \prod_{t=1}^{T} P(y_t \mid y_{<t}, x) \tag{2}$$

where y_t is the token generated at timestep t, and $y_{<t}$ represents all previously generated tokens. The model iteratively computes the probability distribution over possible next tokens, incorporating both the input context and the sequence of preceding tokens to ensure coherence and relevance.

At each timestep t, the model applies a causal self-attention mechanism (Vaswani et al., 2017), restricting attention to previously seen tokens. This prevents information leakage from future tokens and ensures that the model generates answers in a left-to-right fashion, consistent with autoregressive decoding.

Input - Botanical Description (Context)
Each pinna has (5-)7-13 pairs of opposite, sessile, leathery, subtrapezoidal or oblong, sometimes slightly arched leaflets. The blade is glabrous on both sides, with raised venation. The spike-like cluster inflorescences are densely tomentose and pedunculated, measuring 4 to 9 cm long.
Input - Question
How long are the spike-like cluster inflorescences?
Output - Answer
The inflorescences measure 4 to 9 cm long.

Table 3: Example of a training sample for the Answer Generation stage. The model learns to generate an answer from a given context and question.

4.3 Low-Rank Adaptation (LoRA)

To efficiently adapt LLaMA models to the botanical domain, we employ Low-Rank Adaptation (LoRA) with a rank of 64 and a scaling factor of 16. The rank parameter defines the dimensionality of the low-rank update matrices, balancing the parameter efficiency with the ability to capture domain-specific adaptations. The scaling factor controls how much LoRA-modified parameters contribute to weight updates, which prevents excessive deviation from the original model while still enabling domain adaptation.

We apply LoRA modifications to the self-attention layers (query, key, value, and output projections) and the feedforward network layers (gate, up, and down projections), as these components primarily govern contextual representation learning. By fine-tuning these layers while keeping others frozen, we enable F-LoRA-QA to effectively learn botanical domain-specific patterns while retaining the general linguistic knowledge embedded in the base LLaMA model without introducing excessive computational overhead.

To optimize memory usage, we enable gradient checkpointing, which reduces memory consumption during backpropagation and makes fine-tuning feasible on limited hardware resources (an NVIDIA A100 GPU provided by Google Colab). Additionally, we employ AdamW (Loshchilov and Hutter, 2019) optimization with fused gradients to improve training stability and accelerate convergence, ensuring effective adaptation to the botanical QA task while preserving generalization capabilities.

Due to hardware constraints (a single NVIDIA A100 GPU), fine-tuning LLaMA 3.3 70B was infeasible, as it requires multiple high-memory GPUs. LLaMA 2-7B and LLaMA 3-8B were chosen as practical alternatives to balance performance and resource efficiency.

5 Evaluation

To assess the effectiveness of F-LoRA-QA in botanical question-answering, we conduct both automatic and expert-based evaluations. Automatic evaluation quantifies question and answer quality using standard NLP metrics (ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2020)), while expert evaluation assesses five qualitative aspects: Accuracy, Completeness, Relevance, Fluency, and Botanical Terminology Usage.

5.1 Automatic Evaluation

We evaluate F-LoRA-QA by comparing its performance against the base LLaMA models (LLaMA 2-7B and LLaMA 3-8B) without fine-tuning.

We report ROUGE-1, ROUGE-2, ROUGE-L, BLEU, BERTScore, and Exact Match scores for both base and fine-tuned models, including Precision, Recall, and F1 for each. ROUGE and BLEU measure lexical similarity, BERTScore captures semantic similarity, and Exact Match reflects strict answer correctness. We include ROUGE-1, ROUGE-2, and ROUGE-L to capture different granulari-

Metric	Base LLaMA-2	Base LLaMA-3	Fine-tuned LLaMA-2	Fine-tuned LLaMA-3
BLEU	10.99%	2.18%	44.51%	**46.94%**
ROUGE-1 Precision	48.28%	6.75%	76.04%	**79.26%**
ROUGE-1 Recall	51.03%	7.28%	81.84%	**82.26%**
ROUGE-1 F1	44.29%	5.78%	75.37%	**78.23%**
ROUGE-2 Precision	31.2%	4.86%	65.3%	**69.17%**
ROUGE-2 Recall	34.02%	5.19%	70.12%	**71.89%**
ROUGE-2 F1	29.02%	4.23%	64.48%	**68.13%**
ROUGE-L Precision	44.85%	6.59%	73.63%	**77.03%**
ROUGE-L Recall	47.02%	6.96%	79.14%	**79.98%**
ROUGE-L F1	40.97%	5.6%	72.98%	**76.08%**
BERTScore Precision	78.57%	67.31%	91.3%	**92.5%**
BERTScore Recall	79.08%	59.83%	92.64%	**92.83%**
BERTScore F1	78.58%	63.18%	91.8%	**92.54%**
Exact Match	2.01%	0.24%	23.57%	**25.04%**

Table 4: Automatic evaluation results comparing base and fine-tuned LLaMA models on the F-LoRA-QA dataset.

Metrics	Definition
Accuracy	Are the answers factually correct based on the provided context?
Completeness	Do the answers sufficiently address all aspects of the question?
Relevance	Are the answers focused and free of irrelevant details?
Fluency	Are the answers grammatically correct and well-structured?
Botanical Terminology Usage	Are domain-specific terms used correctly?

Table 5: Expert-based evaluation criteria, including their definitions and their scoring system.

ties: unigram overlap (content recall), bigram co-occurrence (local fluency), and longest common subsequence (sentence-level structure).

Although originally designed for summarization and translation, these metrics are widely used in question generation (Du et al., 2017). Since question quality often extends beyond surface-level overlap, BERTScore provides a more flexible evaluation of semantic alignment when wording varies.

Table 4 presents the performance comparison between the base and fine-tuned models across multiple metrics.

5.2 Expert-Based Evaluation

To assess the quality of generated answers beyond automated metrics, we conducted an expert-based evaluation. We selected 100 random samples from the evaluation dataset for expert assessment, focusing on responses generated by the fine-tuned LLaMA 3-8B model, as it achieved the highest performance in automatic evaluation.

The biodiversity expert rated each response on a 1-5 Likert scale across five key aspects: Accuracy, Completeness, Relevance, Fluency, and Botanical Terminology Usage, as defined in Table 5.

Table 6 presents the average scores, indicating that the model performs consistently well across all aspects, with particularly high ratings in Accuracy (4.74) and Botanical Terminology Usage (4.78).

Evaluation Metrics	Average Score
Accuracy	4.74 / 5
Completeness	4.53 / 5
Relevance	4.48 / 5
Fluency	4.48 / 5
Botanical Terminology Usage	4.78 / 5

Table 6: Average expert evaluation scores for responses generated by the fine-tuned LLaMA 3-8B model.

6 Discussion

Exact Match scores remain relatively low, likely reflecting the limitations of strict string matching in a domain where multiple valid paraphrases can exist. As our reference answers, though verified, are not exhaustive, Exact Match likely underestimates correctness.

In contrast, consistently high BERTScores across fine-tuned models indicate strong semantic alignment with reference answers. Improvements in BLEU, ROUGE, and Exact Match after fine-

tuning confirm that **F-LoRA-QA** improves both lexical precision and response relevance.

Since automatic metrics offer limited information on output quality, we conducted expert evaluations covering accuracy, completeness, fluency, and botanical terminology. This provided a more comprehensive validation of the model's outputs.

Although LoRA and instruction tuning are established techniques, their application to French botanical QA is, to our knowledge, novel. **F-LoRA-QA** shows that domain-specific LLM adaptation can support structured trait extraction from floristic texts, addressing a key gap in biodiversity NLP.

Future directions include exploring encoder-only models (e.g., CamemBERT (Martin et al., 2020b)) for trait classification or question filtering, as well as multilingual extension and integration with symbolic reasoning to enhance robustness and interpretability.

Expert evaluation shows competitive performance, with average scores ranging from 4.48 to 4.78 in all metrics (Table 6). Accuracy, completeness, and terminology usage are the model's strengths, while fluency and relevance show room for improvement.

For example, the phrase ***"of this plant"*** in ***"The flowers of this plant are white"*** is redundant and reduces fluency. In terms of relevance, the model sometimes overgenerates information, as in: ***"The reddish style surmounts the 3-celled ovary"*** in response to the question ***"What is the color of the style located at the top of the ovary?"***, where the question asked only for color. In contrast, when questions require multiple details, the model may omit parts of the answer, leading to incompleteness.

These observations highlight opportunities to refine the answer generation strategy.

7 Conclusion

We introduced **F-LoRA-QA**, a two-stage pipeline for French botanical question-answering that fine-tunes LLaMA models using LoRA for efficient domain adaptation. By combining question and answer generation, the system improves information extraction from botanical texts.

Fine-tuning produced notable improvements across BLEU, ROUGE, BERTScore, and Exact Match, with expert evaluation confirming strong performance (4.48–4.78/5), particularly in accuracy and terminology usage.

Although the model is well-adapted for botanical QA, further improvements in fluency are needed. Future directions include instruction tuning, reinforcement learning from human feedback (RLHF), and integrating structured botanical knowledge to enhance answer quality.

Beyond QA, our approach enables systematic taxonomic comparison: Applying a standardized set of questions across species descriptions can surface morphological similarities, thereby offering a scalable method for species identification and evolutionary analysis. This complements traditional biodiversity NLP by introducing automated trait-level comparison.

Limitations

F-LoRA-QA shows strong performance in botanical QA, but a few limitations remain. First, although the training data include texts from multiple floristic regions, the generalization of the model to unseen taxonomic structures or underrepresented traits remains untested. Second, despite high expert-rated accuracy and terminology use, a few answers lack fluency or completeness. The use of QA pairs generated by GPT-3.5, even with human verification, may introduce stylistic biases. Computational constraints limited fine-tuning to LLaMA 2-7B and 3-8B, leaving larger models (e.g., LLaMA 3-70B) unexplored. Finally, the model does not leverage structured botanical resources such as TAXREF or GBIF, which could improve factual grounding and trait comparison.

Addressing these limitations will enhance F-LoRA-QA's robustness and applicability in biodiversity research.

Acknowledgments

This work is co-funded by the European Union's Horizon Europe research and innovation program Cofund SOUND.AI under the Marie Sklodowska-Curie Grant Agreement No 101081674. This research is also part of the e-COL+ project (ANR-21-ESRE-0053).

References

Nora Abdelmageed, Felicitas Löffler, Leila Feddoul, Alsayed Algergawy, Sheeba Samuel, Jitendra Gaikwad, Anahita Kazem, and Birgitta König-Ries. 2022. Biodivnere: Gold standard corpora for named entity recognition and relation extraction in the biodiversity domain. *Biodiversity Data Journal*, 10:e89481.

Md Adnan Arefeen, Biplob Debnath, and Srimat Chakradhar. 2024. Leancontext: Cost-efficient domain-specific question answering using llms. *Natural Language Processing Journal*, 7:100065.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text.

Aman Berhe, Guillaume Draznieks, Vincent Martenot, Valentin Masdeu, Lucas Davy, and Jean-Daniel Zucker. 2023. AliBERT: A pre-trained language model for French biomedical text. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 223–236, Toronto, Canada. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension.

Ryan A. Folk, Robert P. Guralnick, and Raphael T. LaFrance. 2024. Floratraiter: Automated parsing of traits from descriptive biodiversity literature. *Applications in Plant Sciences*, 12(1):e11563.

Roselyn Gabud, Portia Lapitan, Vladimir Mariano, Eduardo Mendoza, Nelson Pampolina, Maria Art Antonette Clariño, and Riza Batista-Navarro. 2023. A hybrid of rule-based and transformer-based approaches for relation extraction in biodiversity literature. In *Proceedings of the 2nd Workshop on Pattern-based Approaches to NLP in the Age of Deep Learning*, pages 103–113, Singapore. Association for Computational Linguistics.

Andrew V. Gougherty and Hannah L. Clipp. 2024. Testing the reliability of an ai-based large language model to extract ecological information from the scientific literature. *npj Biodiversity*, 3(1):13.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Adeline Kerner, Elie Mario Saliba, Sylvain Bouquin, Rémy Portier, and Régine Vignes-Lebbe. 2025. An xper3 reference guide for taxonomists: a collaborative system for identification keys and descriptive data. *European Journal of Taxonomy*, 987(1):281–302.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. BioASQ-QA: A manually curated corpus for Biomedical Question Answering. *Scientific Data*, 10(1):170.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Nicolas Le Guillarme and Wilfried Thuiller. 2022. Taxonerd: Deep neural models for the recognition of taxonomic entities in the ecological and evolutionary literature. *Methods in Ecology and Evolution*, 13(3):625–641.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Diego Marcos, Robert van de Vlasakker, Ioannis N. Athanasiadis, Pierre Bonnet, Hervé Goëau, Alexis Joly, W. Daniel Kissling, César Leblanc, André S. J. van Proosdij, and Konstantinos P. Panousis. 2025. Fully automatic extraction of morphological traits from the web: Utopia or reality? *Applications in Plant Sciences*, 13(3):e70005.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020a. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020b. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Fabricio De Jesus Rios Montero, Ervin Rodríguez, and Maria Mora Cross. 2024. Relation extraction from unstructured species descriptions using taxonnerd and llama 2 7b. *Biodiversity Information Science and Standards*, 8:e142382.

Ayoub Nainia, Régine Vignes Lebbe, Eric Chenin, Maya Sahraoui, Hajar Mousannif, and Jihad Zahir. 2024a. Floraner: a named entity recognition dataset for botanical french text. Dataset available on Zenodo.

Ayoub Nainia, Régine Vignes-Lebbe, Eric Chenin, Maya Sahraoui, Hajar Mousannif, and Jihad Zahir. 2024b. Floraner: A new dataset for species and morphological terms named entity recognition in french botanical text. *Data in Brief*, 56:110824.

Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020. Semantic graphs for generating deep questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1463–1475, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Qiming Wang and Raul Castro Fernandez. 2024. Fabricqa-extractor: A question answering system to extract information from documents using natural language questions.

Xiaowei Wang, Mingdan Zhang, Hao Liu, Xiaodong Ma, Yingchao Liu, and Yitong Chen. 2024. Chatbbnj: a question–answering system for acquiring knowledge on biodiversity beyond national jurisdiction. *Frontiers in Marine Science*, 11.

Xianjun Yang, Junfeng Gao, Wenxin Xue, and Erik Alexandersson. 2024. Pllama: An open-source large language model for plant science.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.