# 微软放大招：基于RAG与Fine-Tuning的数据整合策略探索



> RAG VS FINE-TUNING: PIPELINES, TRADEOFFS, AND A CASE STUDY ON AGRICULTURE

本文探讨了两种常见的方法：Retrieval-Augmented Genera-tion（RAG）和Fine-Tuning，用于将专有和领域特定数据整合到大型语言模型中。作者提出了一个包括提取信息、生成问题和答案、使用它们进行Fine-Tuning以及利用GPT-4评估结果等阶段的管道，并针对多个流行的LLM进行了分析。作者还设计了一些指标来评估不同阶段的性能，并在农业领域进行了深入的研究。结果显示，该管道能够有效地捕捉地理特定知识，并证明了RAG和Fine-Tuning的效果。最终，这些结果为将LLM应用于其他工业领域铺平了道路。

## 论文方法

## 方法描述

该论文提出了一种基于人工智能技术的农业领域知识问答系统。该系统由五个主要组成部分构成：数据采集、信息提取、问题生成、答案生成和模型优化。其中，数据采集阶段通过收集来自美国、巴西和印度等国家的相关文档来构建知识库；信息提取阶段利用自然语言处理技术将文本转化为结构化数据；问题生成阶段使用Guidance框架生成高质量的问题；答案生成阶段采用Retrieval-

Augmented Generation（RAG）方法结合预训练的语言模型生成准确的答案；最后，在模型优化阶段采用了低秩适应（LoRA）算法对模型进行微调以提高其性能。

## 方法改进

与传统的基于规则或模板匹配的知识问答系统相比，该系统具有更高的灵活性和准确性。同时，该系统也克服了传统系统的局限性，例如无法应对复杂的问题或语义歧义等问题。此外，该系统还采用了多种先进的自然语言处理技术和深度学习模型，如Guidance框架、RAG方法和LoRA算法等，从而进一步提高了系统的性能和效率。

## 解决的问题

该系统解决了农业领域知识问答中存在的诸多问题，包括缺乏高质量的数据集、难以处理复杂的语义关系以及需要大量的手动工作等。通过该系统，用户可以方便地获取有关农业领域的各种信息，并获得准确和可靠的答案。这有助于提高农业生产效率、降低成本并促进可持续发展。

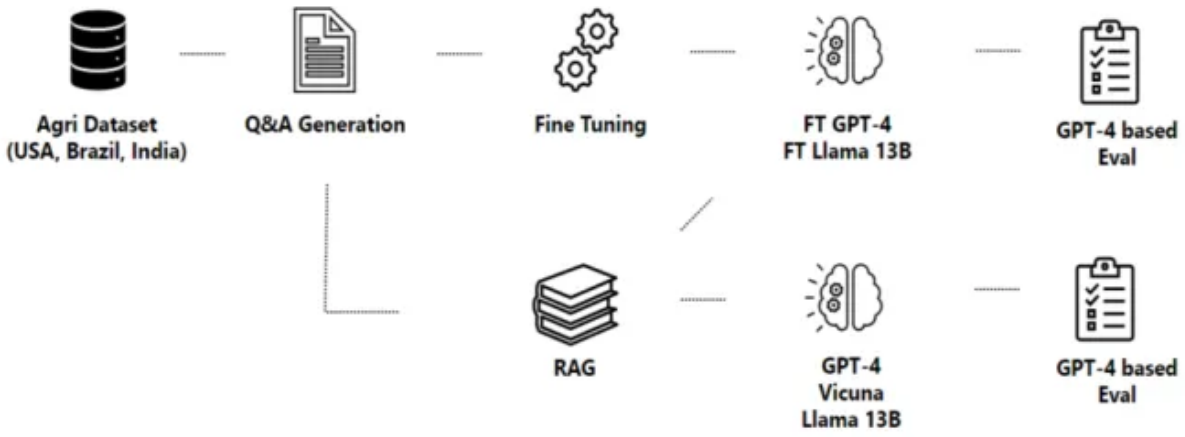


Figure 1: Methodology pipeline. Domain-specific datasets are collected, and the content and structure of the documents are extracted. This information is then fed to the Q&A generation step. Synthesized question-answer pairs are used to fine-tune the LLMs. Models are evaluated with and without RAG under different GPT-4-based metrics.

## 论文实验

本文主要介绍了使用大型语言模型（LLM）生成问答对的质量评价和比较实验。作者通过多个实验探究了不同模型和设置下的表现，并提出了改进方法。

首先，作者进行了三个大型语言模型（GPT-3、GPT-3.5和GPT-4）的问答质量实验，使用不同的上下文环境来评估它们的表现。他们使用了一系列度量标准，如相关性、全球相关性、覆盖范围、重叠度、多样性、细节和流畅性等，来评估这些模型在生成问答对方面的质量。作者还探讨了各种度量方式如何影响对问答对质量的评估，并提供了详细的例子和解释。

接下来，作者进行了几个对比实验：

1. Q&A Quality 实验：在这个实验中，作者比较了三种大型语言模型（GPT-3、GPT-3.5和GPT-4）在不同上下文环境下的问答质量。他们使用了多个度量标准来评估这些模型的表现，并提供了详细的分数和解释。
2. Context Study 实验：这个实验研究了不同上下文环境下大型语言模型的表现。作者将问题分为三种类型：无上下文、有上下文和外部上下文，并使用多个度量标准来评估这些模型的表现。
3. Model to Metrics Calculation 实验：在这个实验中，作者比较了GPT-3.5和GPT-4在计算度量标准时的行为。他们使用了相同的问答对，并观察到GPT-4在覆盖率方面表现更好，但在多样性和重叠度方面表现较差。
4. Combined vs Separated Generation 实验：这个实验研究了同时生成问题和答案与单独生成问题和答案之间的效率差异。
5. Retrieval Ablation Study 实验：这个实验评估了增强型知识的语言模型（RAG）在提供额外上下文信息以帮助回答问题时的表现。
6. Fine-tuning 实验：这个实验比较了微调和基础指导训练的模型在学习新知识方面的性能。

总的来说，本文通过一系列实验比较了不同模型和设置下的表现，并提出了改进方法。这为使用大型语言模型生成问答对的研究提供了有价值的参考。

| **Question** | **Global Relev. Score ↑** | **Explanation** |
|---|---|---|
| How can diversifying or intensifying cropping systems benefit growers in the inland Pacific Northwest? | 5 | This question is highly informative and is likely to be asked by a farmer in the inland Pacific Northwest region who wants to optimize their crop yields and profitability. By asking this question, the farmer wants to understand how diversifying or intensifying their cropping systems can benefit them in the long run. |
| What is the purpose of downscaling global climate models (GCMs) in regional studies? | 1 | While climate change affects farming, this question is too technical in nature and not directly related to farming activities. It is unlikely that a farmer would ask this question. |

Table 3: Examples of Global Relevance scores 1 and 5 for questions generated and present in the WA state benchmark dataset.

| Question | Coverage Score ↑ | Explanation |
|---|---|---|
| What are the potential impacts of climate change on dryland wheat production systems in the inland Pacific Northwest? | 5 | The context provides specific information on climate change impacts on dryland wheat production systems in the inland Pacific Northwest, including temperature-induced drought conditions and increased frequency and severity of extreme weather events. The context also notes how climate change may exacerbate historical conditions linked to soil erosion events in the region. |
| How does the topography of the Pacific Northwest influence the region's climate and agricultural zones? | 1 | The context provides information about wheat production, climate change, soil health, tillage systems, crop residue management, rotational diversification, soil fertility, soil amendments, precision agriculture, weed management, and disease management in the region, but it does not directly address the influence of the topography on the climate and agricultural zones. |

Table 4: Examples of Coverage scores 1 and 5 for questions generated and present in the WA state benchmark dataset.

| Question | Fluency Score ↑ | Explanation |
|---|---|---|
| How do wind and water erosion impact the cereal-producing regions of the Inland Pacific Northwest? | 5 | The question is fluent, coherent, and makes sense. It is a clear and specific question that is directly related to the concerns and interests of farmers who produce cereals in the Inland Pacific Northwest. It demonstrates an understanding of the potential threats posed by wind and water erosion to crop production and suggests a desire to learn more about these issues in order to mitigate their impact on agricultural productivity. |
| Why has the reliability of the Haney's test been questioned? | 3 | The question is coherent and makes sense, but it lacks context. It is unclear what Haney's test refers to and what the question is trying to achieve. |

Table 5: Examples of Fluency scores 3 and 5 for questions generated and present in the WA state benchmark dataset. One should note that almost no values below 3 were obtained.

| Question | Relevance Score ↑ | Answer |
|---|---|---|
| What is the purpose of collecting soil compaction data? | 5 | The purpose of collecting soil compaction data is to determine the extent of soil compaction, which can help farmers make informed decisions about soil management practices and improve long-term soil health. |
| What is the only species of spider documented as venomous to humans in the Pacific Northwest? | 1 | Based on the information provided, the only species of spider documented as venomous to humans in the Pacific Northwest is the Pacific Rattlesnake, *Crotalus oreganus oreganus*. |

Table 7: Examples of Relevance scores 1 and 5 for answers generated using a fine-tuned `Llama2-13b model`.

| Succinctness ↑ | |
|---|---|
| **Score** | **5** |
| Question | What is the scientific name of the white pine weevil? |
| Answer | The scientific name of the white pine weevil is Pissodes strobi. |
| Reference Answer | Pissodes strobi |
| Explanation | The response is on point and does not contain any additional information. The user is able to understand the point quickly. |
| **Score** | **2** |
| Question | What is the purpose of salting fish before smoking? |
| Answer | The purpose of salting fish before smoking is to help remove moisture from the fish and to preserve it by inhibiting the growth of bacteria. Salt draws out the moisture from the surface of the fish, making it firm and creating a barrier that prevents bacteria from multiplying. It also enhances the flavor of the fish and helps to prevent it from spoiling. This process is commonly used in curing and preserving fish, and is an important step in the preparation of smoked fish. |
| Reference Answer | Salting fish before smoking preserves the fish by reducing its moisture content. |
| Explanation | The predicted answer provides a lot of additional information about the process of salting fish, including how it works and why it's important. While this information is relevant and informative, it goes beyond the scope of the original question, which simply asked for the purpose of salting fish before smoking. Therefore, the response is verbose. |

Table 9: Examples of Succinctness scores for GPT-4 generated answers.

## 论文总结

## 文章优点

该论文针对大型语言模型在农业领域的应用进行了全面的研究，并提出了一个综合的LLM管道来生成高质量、行业特定的问题和答案。具体来说，本文的主要优点包括：

1. 全面评估了各种大型语言模型在回答农业相关问题上的性能表现，为后续研究提供了重要的参考。
2. 研究了两种有效的技术：Retrieval-Augmented Generation（RAG）和fine-tuning，以提高大型语言模型的性能。
3. 提出了用于评估问题和答案质量的一系列指标，这有助于更准确地衡量模型的表现。
4. 在实验中，使用基准数据集对多个模型进行了全面测试，结果表明GPT-4具有更好的性能，但其成本需要考虑。

## 方法创新点

本文的方法创新点主要体现在以下几个方面：

1. 创新性地提出了一种LLM管道，可以生成高质量、行业特定的问题和答案，以满足不同行业的实际需求。
2. 通过比较不同的技术和模型，得出了使用RAG和fine-tuning的有效策略，这对进一步改进模型性能具有重要意义。
3. 使用一系列自定义的指标来评估问题和答案的质量，这对于更准确地衡量模型表现非常重要。

## 未来展望

基于本研究的结果，未来可以从以下几个方向继续探索：

1. 进一步优化LLM管道，以更好地适应不同行业的实际需求。
2. 对其他领域的大规模语言模型进行类似的研究，以便将这些方法扩展到更多应用场景。
3. 研究如何利用多模态信息（如图像和文本）来进一步提升模型性能，特别是在涉及视觉信息的场景下。
4. 探索如何在更大范围内收集和整理与特定行业相关的知识，以进一步丰富LLM的知识库。