LEVERAGING MULTIMODAL LLMS FOR PLANT SPECIES IDENTIFICATION AND EDUCATIONAL INSIGHTS

Yuze Du
Northeast Yucai Foreign Language School
Benxi, Liaoning, China
robinhan1999@gmail.com

Yingjia Wang
Northeast Yucai Foreign Language School
Shenyang, Liaoning, China
cyann417@163.com

Eric Zhao
Department of Applied Math
Northwestern University
Evanston, Illinois, USA
yingzhao2018@u.northwestern.edu

ABSTRACT

In this study, we investigate the potential of multimodal large language models (LLMs) for plant species identification and educational enrichment. Using an annotated dataset focused on fungi, particularly those classified as edible or non-edible, we implement a practical application that allows users to upload plant images. The LLM then identifies the species, determines its edibility, and provides detailed information on its characteristics. For edible species, the model offers culinary insights and preparation methods, while also delivering comprehensive educational content on plant ecology and cultural significance. Our approach showcases the ability of LLMs to bridge image recognition with rich, text-based knowledge, facilitating an interactive learning experience that promotes plant literacy and practical understanding. This study highlights the effectiveness of LLMs in educational tools and their potential to enhance public awareness of plant species, including fungi, through visual and contextual data fusion.

1 Introduction and Background

The rapid advancement of large language models (LLMs), such as GPT-3, BERT, and their successors, has revolutionized natural language processing (NLP) by enabling machines to understand and generate human language with impressive accuracy. [1] More recently, the development of multimodal LLMs—models capable of processing both text and images—has expanded the possibilities of AI applications. These models, such as CLIP and GPT-4's multimodal capabilities, bridge the gap between textual and visual information, making them well-suited for tasks involving image recognition and text generation. Leveraging multimodal LLMs can significantly enhance a variety of applications, particularly those requiring contextual understanding from visual data, such as species identification from plant images. [2]

Identifying plants through photographs has long been a necessity for both professional botanists and amateur naturalists. [3] The accurate identification of plant species can help in research, conservation efforts, and even casual foraging. However, existing tools often rely on simple image classification algorithms that lack depth in providing educational content. For general users, these tools may not offer enough context, such as how the identified plant fits into its ecological environment, its role in biodiversity, or whether it poses any dangers—such as toxicity or allergens. This lack of information is a significant gap, especially considering the increasing interest in sustainable living, foraging, and understanding our natural surroundings.

There is a growing need for educational tools that not only identify plants but also provide comprehensive insights, including potential ecological benefits, protection statuses, and safety precautions. Many plants and fungi are misidentified, which can lead to harmful consequences. For example, mistaking toxic mushrooms for edible varieties can have dire health outcomes, as mushroom poisoning is not uncommon in some regions. Educating the general public about

Figure 1: Summary statistics and examples of the fungi dataset

the environment, as well as the risks and benefits of various species, is essential to promote responsible interaction with nature. [4] Moreover, understanding plant ecology and recognizing rare or endangered species can aid in conservation efforts and protection of biodiversity.

LLMs, particularly multimodal ones, are well-suited to address these challenges. Unlike traditional image classifiers that are limited to visual recognition, LLMs can combine their contextual understanding of text with image analysis to generate rich educational content. These models can provide species identification from a single photo, followed by detailed information about the plant's edibility, toxicity, ecological role, and even practical uses such as medicinal benefits or culinary preparation. This approach not only improves the accuracy of identification but also adds educational value by explaining the significance of the species, thereby bridging the gap between image-based recognition and text-based knowledge. Moreover, LLMs require no additional training once implemented, simplifying their use in various applications.

In this study, we focus on an important use case: the identification of edible fungi and their corresponding educational insights. In many regions, such as China, mushroom poisoning is a significant public health issue, primarily caused by the misidentification of poisonous mushrooms as edible species. According to recent research [5], mushroom poisonings often result in gastrointestinal distress, but more severe cases can lead to liver failure, kidney damage, and even death. From 2010 to 2022, over 10,000 mushroom poisoning outbreaks were reported in China, resulting in nearly 800 deaths. [5] These outbreaks peak between May and October, a time when foraging activity is highest. The need for a reliable identification tool that can prevent mushroom poisoning is evident.

We present an application that leverages multimodal LLMs to identify fungi species and determine their edibility based on user-uploaded images. This tool goes beyond mere identification by offering educational content on the risks associated with consuming certain fungi, cooking methods for edible varieties, and ecological information to help users better understand their environment. Given the high stakes associated with fungi misidentification, especially in regions with a tradition of foraging, such a tool has the potential to save lives by reducing accidental poisonings. This study showcases the utility of LLMs in practical applications while demonstrating their ability to educate the public on important ecological and safety considerations surrounding fungi and plants.

2 Methods

2.1 Development of the Application

We developed a web-based application that allows users to upload an image and receive educational insights about the plant or fungus species in the image. The app was built using the OpenAI GPT-4 API, integrated with the Streamlit framework for ease of use. When a user uploads an image, it is resized to a resolution of 512x512 pixels. This resizing helps reduce the token usage for the GPT-4o API, improves consistency, and minimizes computational load while retaining the visual details necessary for accurate species identification.

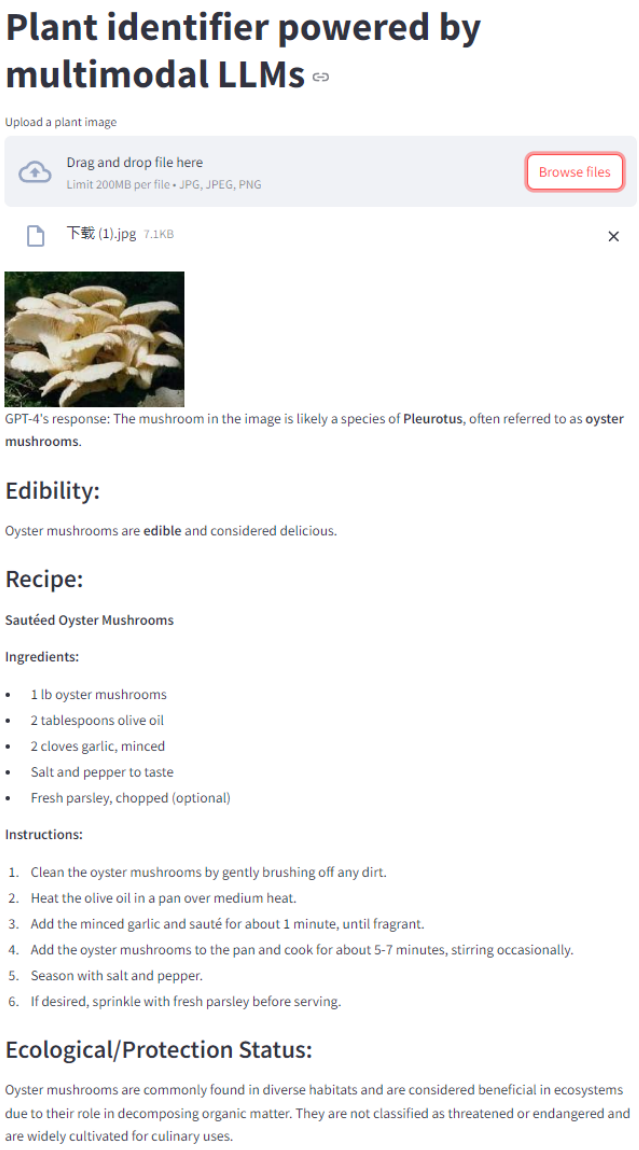

Figure 2: Interface for the plant identifier

The application is hosted on a local workstation, accessible through port access. This enables real-time processing and feedback, ensuring that the image is processed swiftly and insights are delivered promptly to the user. The backend processes include image handling, resizing, and sending data to the OpenAI GPT-4o API for prediction.

2.2 Prompt Engineering

The accuracy of species identification and educational insight generation depends heavily on prompt engineering. First, the model identifies the species shown in the uploaded image, followed by determining whether the plant or fungus is edible. If the species is determined to be poisonous or inedible, the application provides safety information and educational guidance.

If the species is edible, the model suggests appropriate culinary recipes. To support the process, we incorporated a crowdsourced list of common edible fungi. For cases where GPT-4 is unable to identify the species with confidence, the model requests additional images from different angles. The prompt is structured to ensure the model provides

Table 1: Performance of four deep convolutional neural networks and one multimodal LLM on the classification of poisonous mushrooms.

Model	Accuracy	AUC	Precision	Recall	F1
AlexNet	0.700	0.748	0.705	0.750	0.727
VGG16	0.689	0.735	0.689	0.833	0.755
DenseNet121	0.744	0.764	0.756	0.708	0.731
ResNet50	0.756	0.809	0.782	0.750	0.766
GPT-4o	0.773	0.821	0.797	0.762	0.777

species-specific protection and conservation insights, educating users on how the plant or fungus fits into the larger ecological context.

2.3 Test Cases

To evaluate the performance of our application, we curated a dataset consisting of fungi images found in China. We selected 200 images of edible mushrooms and 250 images of poisonous mushrooms. The edible mushroom images were based on the Chinese Edible Fungi Checklist [6], and the poisonous mushroom images were drawn from the checklist revised by Bau et al. [7].

The dataset is balanced to cover a diverse set of mushroom species found in China, ensuring the model encounters a variety of real-world test cases. Each species is represented only once in the dataset, with no duplicate images. This dataset is open-sourced for public access to benefit researchers interested in mushroom classification. Access to the dataset is available at: `https://drive.google.com/drive/folders/13NFDI5UhcLHPSL2WMcOrFs6QhhmrDjxQ?usp=sharing`.

2.4 Evaluation

For the evaluation in your study, the performance of the multimodal LLM is compared with four well-known deep convolutional neural networks: AlexNet, VGG15, DenseNet121, and ResNet50. [8] The key distinction here is that the multimodal LLM is training-free, meaning it does not require specific training on the dataset, whereas the deep CNNs were trained on the training split of the evaluation dataset. This dataset contains approximately 100 images for each class—edible and poisonous mushrooms.

For the deep CNNs, the models were fine-tuned to optimize classification performance based on the mushroom dataset, where both classes (edible and poisonous) were used. The evaluation was performed by comparing accuracy, F1 score, and other metrics, providing insights into the benefits and trade-offs between using training-free approaches like multimodal LLMs and conventional CNNs that require extensive training.

This setup highlights the efficiency and ease of deployment of multimodal LLMs for certain tasks, compared to CNNs, which require significant computational resources and time for training, especially with smaller datasets like the one used in this study.

2.5 Implementation

The entire pipeline was implemented using Python 3.8. Image preprocessing, including resizing, was performed using the Pillow library. Data handling and preprocessing were managed using Pandas and NumPy. The OpenAI GPT-4o API was used for plant and fungus identification, integrated into the Streamlit application for a user-friendly interface.

For model evaluation, the Scikit-learn library was employed to calculate accuracy, AUC, precision, recall, and F1-score. All processes were run on a high-performance local workstation, and visualizations such as confusion matrices and ROC curves were created using Matplotlib and Seaborn.

3 Results

The results of the developed Streamlit interface demonstrate a user-friendly and educational platform that allows users to upload an image of a plant or fungus. After uploading, the image is processed using GPT-4 to identify the species depicted. If the species is recognized, the system provides relevant information such as edibility, protection status, and ecological insights. For edible species, the platform also suggests recipes, as shown in the oyster mushroom

example, which includes step-by-step cooking instructions. If the species remains unidentifiable, the user is prompted to upload more images for better analysis. This combination of species identification and educational content makes the application a valuable tool for enhancing plant and fungi awareness, especially in terms of safe consumption and conservation knowledge. The platform balances simplicity of use with the delivery of rich, informative content, benefiting both casual users and plant enthusiasts.

The performance comparison between the multimodal LLM (GPT-4o) and four deep convolutional neural networks (CNNs) — AlexNet, VGG16, DenseNet121, and ResNet50 — in classifying poisonous mushrooms is summarized in Table 1. The multimodal LLM outperformed the traditional CNN models across all evaluation metrics.

ResNet50, which was the best-performing CNN model, achieved an accuracy of 0.756, with an AUC of 0.809 and an F1 score of 0.766. However, GPT-4o surpassed ResNet50, with an accuracy of 0.773 and an AUC of 0.821, showing a marked improvement. The GPT-4o model also achieved the highest F1 score of 0.777, indicating its robustness in handling this binary classification task.

While the CNN models demonstrated competitive performance, particularly ResNet50 and DenseNet121, the multimodal LLM provided an advantage in classification, despite being training-free. This highlights the potential of LLMs in efficiently classifying visual data without the need for intensive model training, offering significant time savings and computational efficiency. Furthermore, the use of GPT-4o also allows for integration with text-based outputs, enabling it to provide educational information in addition to classification results, which is particularly valuable in real-world applications such as mushroom identification and education.

4 Limitation and Conclusion

Despite the promising results of leveraging multimodal large language models (LLMs) for plant and fungi species identification, several limitations exist in this study. One major limitation is the reliance on predefined datasets and curated crowdsourced images, which may not comprehensively capture the diversity of species, especially in uncontrolled environments. Moreover, the model's reliance on user-uploaded images poses a risk of incorrect predictions if the image quality is poor or ambiguous, as LLMs have inherent limitations when faced with noisy data. Another key limitation is the inability of LLMs to generalize beyond species or datasets that were well-represented in the training phase. Thus, novel species or plants with atypical features may yield incorrect classifications. Furthermore, the performance of LLMs in identifying nuanced differences between closely related species remains to be rigorously tested. The system also depends on the GPT-4 API, which raises concerns about long-term accessibility and scalability, given the reliance on proprietary technologies.

In this study, we explored the use of multimodal LLMs to identify plant and fungi species from user-uploaded images, while providing educational insights, such as ecological information and edibility status. Through practical application in mushroom classification, we demonstrated that LLMs offer comparable performance to deep learning models without requiring intensive training, making them highly suitable for lightweight, user-friendly applications. The proposed approach not only simplifies species identification but also enhances user engagement through contextual knowledge delivery, making it a promising tool for plant education and safety awareness. Future research should focus on enhancing the system's robustness, expanding the training dataset, and addressing the limitations associated with rare or atypical species identification.

5 Author Contribution Statement

All authors discuss and define the research problem. Y.W. extracts and pre-process all the research data. Y.D. and E.Z. implement all deep neural networks, analyze and visualize the results. Y.D. and Y.W. write the manuscript advised by E.Z., who also organizes the manuscript to a better academic standard.

References

[1] Mohaimenul Azam Khan Raiaan, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, 2024.

[2] Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, et al. A comprehensive review of multimodal large language models: Performance and challenges across different tasks. *arXiv preprint arXiv:2408.01319*, 2024.

[3] Jaak Pärtel, Meelis Pärtel, and Jana Wäldchen. Plant image identification application demonstrates high accuracy in northern europe. *AoB Plants*, 13(4):plab050, 2021.

[4] Vernon H Heywood. Conserving plants within and beyond protected areas–still problematic and future uncertain. *Plant Diversity*, 41(2):36–49, 2019.

[5] Weiwei Li, Sara M Pires, Zhitao Liu, Jinjun Liang, Yafang Wang, Wen Chen, Chengwei Liu, Jikai Liu, Haihong Han, Ping Fu, et al. Mushroom poisoning outbreaks—china, 2010–2020. *China CDC weekly*, 3(24):518, 2021.

[6] LiWei Zhou, ZhuLiang Yang, HuaAn Wen, T Bau, TaiHui Li, et al. A revised checklist of edible fungi in china. *Mycosystema*, 29(1):1–21, 2010.

[7] Tolgor Bau, Bao HaiYing, LI Yu, et al. A revised checklist of poisonous mushrooms in china. *Mycosystema*, 33(3):517–548, 2014.

[8] Leiyu Chen, Shaobo Li, Qiang Bai, Jing Yang, Sanlong Jiang, and Yanming Miao. Review of image classification algorithms based on convolutional neural networks. *Remote Sensing*, 13(22):4712, 2021.