CONVERSATIONAL IMAGE RECOGNITION CHATBOT SYSTEM

[1]Dr. G. Prasuna, [2]D. Mounika, [3]G. Pujitha, [4]G. Samuel Kiran, [5]G. Aishwarya

[1]Associate Professor, Dept of Computer Science and Engineering, St. Ann's College of Engineering and Technology, Chirala-523187, India.

[2,3,4,5]U. G Student, Dept of Computer Science and Engineering, St. Ann's College of Engineering and Technology, Chirala-523187, India.

ABSTRACT- ***The Conversational Image Recognition Chatbot is an intelligent web-based application designed to analyze uploaded images and provide accurate responses to user queries using Artificial Intelligence. Traditional image recognition systems generally identify objects but fail to support natural language interaction, making it difficult for users to obtain detailed contextual information about an image. To overcome this limitation, the proposed system integrates advanced vision-language capabilities that enable users to upload an image and ask questions in conversational language, receiving meaningful and context-aware responses. The system is developed using HTML, CSS, JavaScript, and Flask for the user interface and backend, while SQLite is used for user authentication and data management. Google Gemini Vision-Language Model serves as the core AI engine for image understanding and natural language response generation. The system employs image preprocessing, visual feature extraction, and prompt-based conversational reasoning to interpret image content and generate accurate answers. This application can be used in education, accessibility support, smart assistance, healthcare, retail, and content analysis, where users require quick and interactive understanding of visual information. By combining computer vision with conversational artificial intelligence, the proposed chatbot provides an efficient, user-friendly, and scalable solution for intelligent image analysis and human-computer interaction.***

KEYWORDS- ***Conversational Image Recognition, Artificial Intelligence, Computer Vision, Google Gemini, Vision-Language Model(VLM), Image Understanding, Natural Language Processing(NLP), Flask Framework, Human-Computer Interaction.***

INTRODUCTION

Artificial Intelligence(AI) and Computer Vision have become essential technologies for analyzing and understanding digital images. Traditional image recognition systems are capable of detecting objects and generating basic descriptions; however, they often lack the ability to interact with users through natural language. This limitation makes it difficult for users to obtain detailed and context-specific information from images in an intuitive manner.

The proposed Conversational Image Recognition Chatbot overcomes this challenge by allowing users to upload an image and ask questions about its content in natural language.

RELATED WORK:

Recent advancements in Artificial Intelligence, Computer Vision, and Vision-Language Models have significantly improved image understanding and conversational interaction. Traditional image recognition systems mainly focused on object detection and image classification, while modern models such as Google Gemini can analyze images and answer user queries in natural language. Existing research on multimodal AI and image captioning has provided the foundation for developing intelligent image-based applications. Inspired by these studies, the proposed Conversational Image Recognition Chatbot integrates the Google Gemini Vision-Language Model with a Flask-based web application to provide accurate, context-aware, and interactive responses, making image analysis more efficient and user-friendly.

LITERATURE SURVEY

Recent studies in Artificial Intelligence, Computer Vision, and Vision-Language Models have improved image analysis and visual understanding. Traditional image recognition systems mainly focus on object detection and classification, while modern models such as Google Gemini enable users to ask questions about images and receive context-aware responses. These advancements have made image recognition more interactive and accurate. The proposed Conversational Image Recognition Chatbot utilizes these technologies to provide an intelligent, user-friendly, and efficient solution for image-based question answering.

EXISTING METHOD

Existing image recognition systems, such as the CLIP (Contrastive Language-Image Pre-training) model, have shown excellent performance in image classification and image-text matching tasks. CLIP learns the relationship between images and text using large-scale datasets, making it effective for

recognizing visual content. However, the model mainly focuses on matching images with predefined text descriptions and does not provide detailed conversational responses to user queries. It also has limited capability to understand complex contextual questions or maintain interactive communication with users. As a result, users may not receive complete explanations or personalized answers about the uploaded image. In addition, implementing such systems for real time web-based interaction requires additional components for conversational reasoning. These limitations reduce the overall user experience, especially in applications that require natural language interaction. To overcome these drawbacks, the proposed system integrates the Google Gemini Vision-Language Model, which enables context aware image understanding and conversational question answering, providing more accurate, interactive, and user-friendly responses.

PROPOSED METHOD

The proposed Conversational Image Recognition Chatbot is designed to overcome the limitations of traditional image recognition systems by providing intelligent and interactive image through natural language communication. Unlike existing methods that mainly identify objects or match images with predefined text, the proposed system allows users to upload an image and ask questions related to its content. The application is developed using HTML, CSS, JavaScript, Flask, and SQLite, while the Google Gemini Vision-Language Models is used to analyze images and generate context-aware responses. The system understands both the visual content and the user's query, enabling accurate and meaningful conversations. It provides better contextual understanding, improved response quality, and a more user-friendly interface compared to conventional image recognition approaches. The proposed method also supports real-time interactions, making it suitable for educational assistance, accessibility, healthcare, and other intelligent applications. By integrating advanced vision language technology with a web-based platform, the system offers an efficient, scalable, and accurate solution for conversational image understanding and resolves the limitations identified in existing image recognition methods.

SYSTEM ARCHITECTURE

The architecture of the proposed Conversational Image Recognition Chatbot consists of several interconnected modules that work together to provide intelligent image analysis. First, the user logs into the application and uploads an image through the web interface. The uploaded image is securely processed by the Flask backend and

forwarded to the Google Gemini Vision-Language Model. The user then enters a question related to the uploaded image. Gemini analyzes both the image and the text query to understand the visual content and context before generating an accurate response. Finally, the chatbot displays the AI-generated answer on the web interface. SQLite is used to manage user authentication and application data, while HTML, CSS, and JavaScript provide an interactive and user-friendly interface. This architecture enables efficient, real-time, and conversational image understanding.

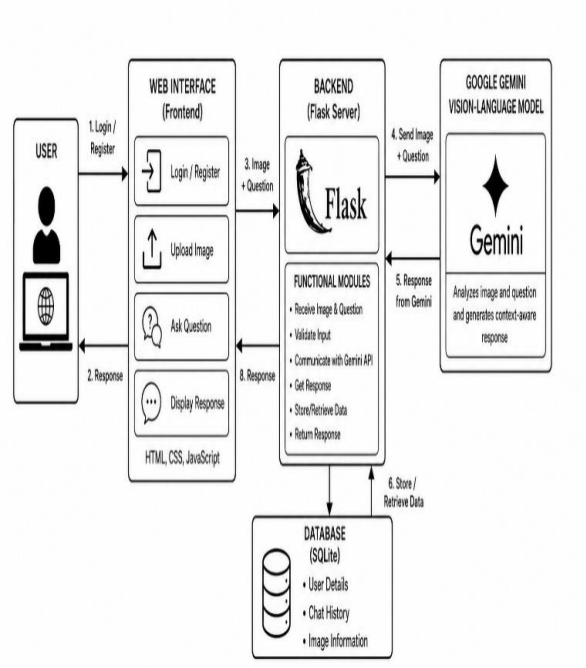

Fig 1: System Architecture

METHODOLOGY DESCRIPTION

User Authentication:

The user first logs into the application using a valid username and password. This ensures that only authorized users can access the chatbot and upload images for analysis.

Image Upload:

After successful login, the user uploads an image through the web interface. The uploaded image is securely transferred to the Flask-backend for further processing.

Question Input:

The user enters a question related to the uploaded image in natural language. This allows the chatbot to understand what specific information the user wants from the image.

Image Processing:

The Flask backend validates the uploaded image and user query before sending them to the Google Gemini Vision-Language Model. This step ensures that valid input is processed efficiently.

AI Analysis:

The Google Gemini model analyzes both the image and the user's question to understand the visual content and context. It then generates an accurate, meaningful, and context-aware responses.

Response Display:

The generated response is sent to the Flask application and displayed on the user interface. The user can view the answer

instantly and continue asking additional questions about the same image.

RESULTS AND DISCUSSION:

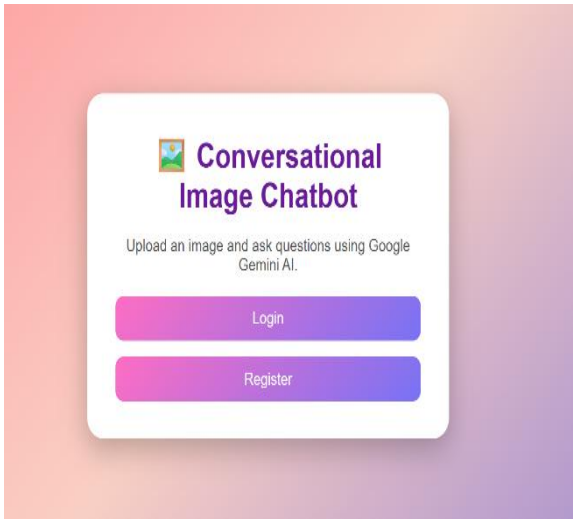

Fig 2: Home page

The home page provides a user-friendly interface where users can upload an image and ask questions related to the image.

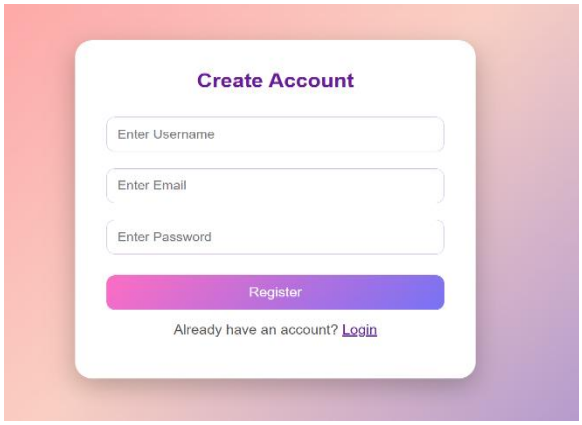

Fig 3: Registration page

The registration page allows users new users to create an account by entering their basic details. It ensures secure user registration before accessing the chatbot features.

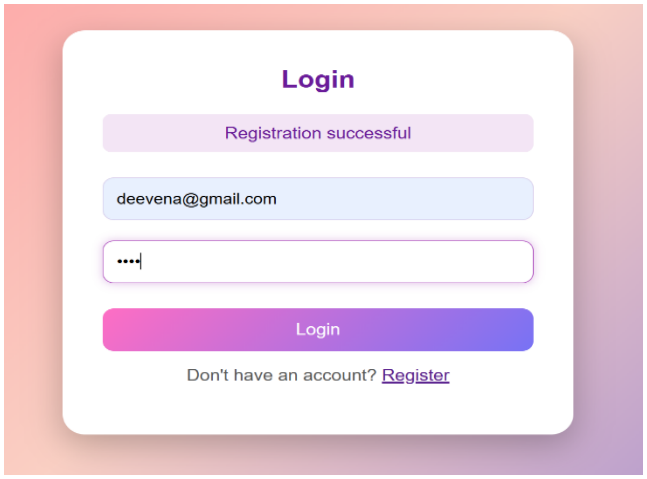

Fig 4: Login page

The login page provides secure access to registered users. The login page process is successfully completed, granting the user access to the Conversational Image Recognition Chatbot.

Fig 5: Image Uploaded

The selected image is uploaded successfully and displayed on the interface for user verification. Once the upload is complete, the image is ready for AI-based analysis and question answering.

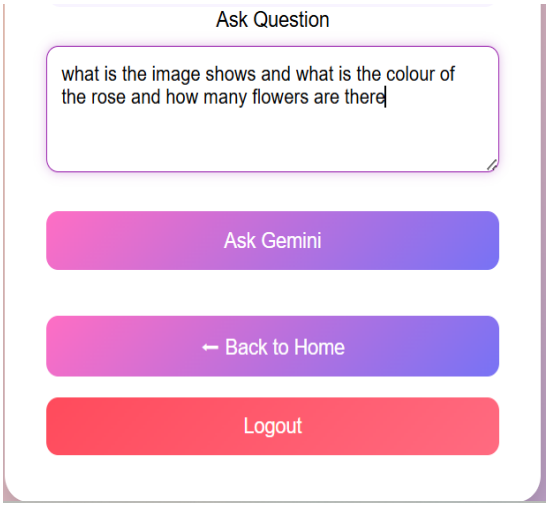

Fig 6: Prompting Question

The user enters a question related to the uploaded image through the prompt box. The system processes the query and prepares it for AI-based image analysis to generate an appropriate response.

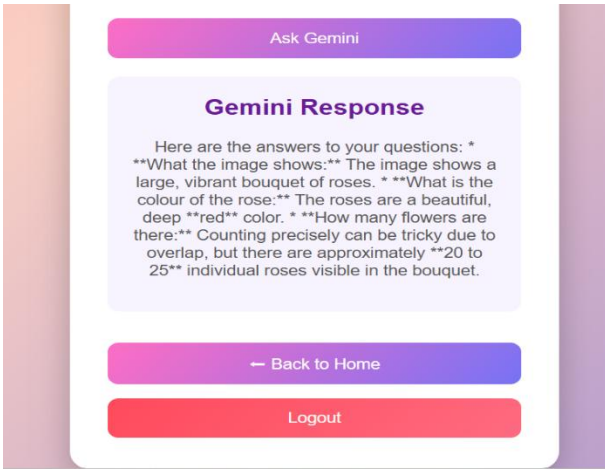

Fig 7: Final Output

The final output displays the AI-generated response based on the uploaded image and the user's query. The chatbot provides meaningful and relevant information, ensuring an efficient and interactive user experience.

CONCLUSION

The proposed Conversational Image Recognition Chatbot successfully integrates Artificial Intelligence, Computer Vision, and the Google Gemini Vision-Language Model to provide intelligent image-based question answering. Unlike traditional image recognition systems, the proposed solution enables users to interact with images through natural language and receive accurate, context-aware responses. The system is developed using HTML, CSS, JavaScript, Flask, and SQLite, making it efficient, user-friendly, and easy to use. It can be applied in various domains such as education, healthcare, accessibility, and smart assistance. Overall, the proposed system improves image understanding, enhances user interaction, and demonstrates the effectiveness of vision-language technology in real-world applications.

FUTURE SCOPE

The proposed Conversational Image Recognition Chatbot can be further

enhanced by integrating multilingual support, allowing users to communication in multiple languages. Future improvements may include voice-based interaction, real-time video analysis, and support for document and medical image understanding. The system can also be integrated with cloud services to improve scalability and performance. In addition, advanced AI models and personalized user assistance can be incorporated to provide more accurate, efficient, and interactive image analysis for a wide range of real-world applications.

REFERENCES

[1] P. Harini, "Sms based wireless e-notice board," *International Journal of Emerging Technology and Advanced Engineering,* vol. 3, no. 3, pp. 181-185, 2013.

[2] P. Harini, "Enhanced packet delivery techniques using crypto-logic riddle on jamming attacks for wireless communication medium," *Int. J. Latest Trends Eng. Technol,* vol. 2, no. 4, pp. 469-478, 2013.

[3] P. Harini, "[Retracted] Machine Learning Algorithms Are Applied in Nanomaterial Properties for Nanosecurity," *Journal of Nanomaterials,* vol. 2022, no. 1, p. 5450826, 2022.

[4] Gemini Team, "Gemini: A Family of Highly Capable Multimodal Models," 2023.

[5] A. Dosovitskiy et al, "An Image is Worth 16*16 Words: Transformers for Image Recognition at Scale," in Proc. Int. conf. Learning Representation(ICLR), 2021.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE Conf. Computer Vision and Pattern and Recognition(CVPR),2016.

[7] O. Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenges," International journal of Computer Vision, vol. 115, no. 3, pp. 21-252, 2015.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Communications of the ACM, vol. 60, no. 6, pp. 84-90, 2017.

[9] C. Szegedy et aI., "Going Deeper with Convolutions," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2015.

[10] T. Chen et al., "A Simple Framework for Contrastive Learning of Visual Representation," in Proc. Int. Conf. Machine Learning(ICML), 2020.

[11] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language

Understanding," in Proc. NAACL-HLT, 2019.

[12] A. Vaswani et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurlPS), 2017.

[13] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2021.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, no. 7553, pp. 436-444, 2015.

[15] B. Bai et al., "Qwen2-VL Technical Report," 2024.