Chatbot to interprets Text and Image

Intern Name: Kaushal Sahu

Internship Duration: 2/7/2025 – 5/5/2025

Organization: NullClass

1. Introduction

This part of the internship focused on extending a traditional text-based chatbot to support multi-modal interactions—specifically, handling both text and image inputs and outputs. The chatbot was built using Google Gemini API and Streamlit, showcasing the integration of visual and textual content in real-time conversations.

2. Background

In modern AI applications, especially in customer service, education, and media, users expect systems to process and generate not just text but also visual content. Multi-modal AI models like Google Gemini offer capabilities to understand and generate responses across both formats, bridging the gap between vision and language.

3. Learning Objectives

- Learn how to integrate Google's Gemini (PaLM) API for multi-modal tasks.
- Enable a chatbot to process user-uploaded images and return contextual responses.
- Build a seamless web-based interface using Streamlit for interaction.
- Understand how to simulate image generation based on prompts

4. Activities and Tasks

- Set up Google Generative AI using the API key and installed google.generativeai.
- Created two core functions:
  - "generate_text_response(prompt)" for text generation.
  - "generate_image_response(prompt)" using placeholder image simulation via picsum.photos.
- Built a Streamlit-based user interface for selecting input types (text/image).
- Integrated image uploading using "st.file_uploader()" and displayed responses accordingly.

- Ensured conditional logic for both input types and safe handling of user inputs.

5. Skills and Competencies

- Technical:
  - API integration with Google Generative AI (Gemini)
  - Streamlit-based UI design for interactive apps
  - Image handling with Python (PIL, requests, io)
  - Working with multi-modal data inputs
- Soft Skills:
  - Adaptability in working with new AI tools
  - Building intuitive user experiences
  - Managing error handling and conditional flows

6. Feedback and Evidence

The chatbot successfully responded to textual prompts(using Gemini AI) as well as Simulated image generation upon user-uploaded image interaction. It further demonstrated smooth interaction in Streamlit without crashes or delays. As it is an API

The evidence or the test responses are been uploaded with a folder named “sample” in the file itself. One can check it before running the actual code. The image generated part is not as perfect as the text because of the API capabilities(And I found no good AI free API to work with the image section). But someone can use different API to work on the image context.

7. Challenges and Solutions

1. Gemini's image generation API access not available directly

SOL- At the very middle of the task I came to know that Gemini’s image generation API was giving to much error, but I have no choice instead of using them. So I simulated it using “https://picsum.photos” to represent image-based responses

2. Keeping the code modular and scalable

SOL- To handle this challenge I had split all the function separately according to there functionality.

8. Outcomes and Impact

- Created a robust base chatbot with dual input capabilities (text and image).
- Demonstrated the potential of using Gemini AI for multi-modal learning environments.
- This base model if modified further can give follow-up answer in any image based prompt or give the response by reding the environment in the image itself.

9. Conclusion

This project highlighted the growing relevance of multi-modal AI systems and how tools like Google Gemini can be effectively leveraged in practical chatbot applications. With a clean Streamlit UI and clear modular logic, the chatbot lays the foundation for more advanced use cases involving image understanding, generation, and interaction. The internship provided valuable insights into modern AI capabilities and practical deployment.