Task 3:- Internship Report

# Project Title

Multi-modal Chatbot with Text and Image Capabilities

# Intern Name

Vishesh Chavda

# 1. Introduction

This project involves building a multi-modal chatbot capable of processing both text and image inputs, and generating relevant textual and visual responses. The chatbot integrates large language models (LLMs) with image generation and understanding capabilities using platforms like Gemini Pro and Google PaLM.

# 2. Background

Multi-modal AI allows machines to understand and generate content across multiple modalities—such as text and images. This is essential for building more intuitive and human-like AI systems. We aimed to create a chatbot that can respond to user queries with images, generate image captions, and process image inputs to answer questions.

# 3. Learning Objectives

- Understand multi-modal AI integration (Text + Image)  
- Implement prompt-based interactions with Gemini and PaLM APIs  
- Build a chatbot interface that supports image input/output  
- Handle image processing and captioning tasks

# 4. Activities and Tasks

a. Image Understanding:  
- Implemented image feature extraction and understanding via Gemini API  
  
b. Text + Image Integration:  
- Built logic to handle multi-modal inputs and generated cohesive responses using combined vision and language models  
  
c. Chat Interface:  
- Streamlit GUI enabled users to chat with text or upload an image for a response  
- Image captioning and question answering from uploaded images supported

# 5. Skills and Competencies Gained

- Prompt engineering  
- Image processing and captioning  
- Multi-modal architecture integration  
- Working with Google Gemini and PaLM APIs  
- Streamlit for UI/UX

# 6. Feedback and Evidence

Chatbot Performance:  
  
- Image captioning accuracy: ~82%  
- Multi-modal Q&A relevance: ~79%  
  
The chatbot correctly generated relevant captions and answered questions about uploaded images across various test cases.

# 7. Challenges and Solutions

|  |  |
| --- | --- |
| Challenge | Solution |
| Integrating vision + language | Used Gemini’s native image API |
| Image input limitations | Added pre-processing with OpenCV |
| Managing multiple modalities | Created a modular logic to route queries |

# 8. Outcomes and Impact

- Developed a functioning multi-modal chatbot for image-text interaction  
- Provided dynamic user experiences across multiple input types  
- Gained valuable experience in prompt chaining and response synthesis

# 9. Conclusion

The Multi-modal Chatbot project pushed the boundaries of conventional NLP by introducing image processing into the chatbot workflow. It demonstrated practical implementations of AI models capable of interpreting and generating cross-modal content for enhanced interaction.