VisionPulse(दिव्य दृष्टि)

AI–Enhanced Learning and Interaction Assistant for Visually Impaired people at IIIT Delhi

M. Mittal 2021538 | **U. Venaik** 2021570 | **A. Kushwaha** 2021514 | **L. Kumar** 2021061 | **T. Singh** 2021569 | **S. Kabra** 2021563

Abstract

"VisionPulse" - our revolutionary idea in the educational domain to cater to the needs of visually impaired students at IIIT Delhi by leveraging multiple facets of Computer Vision and Natural Language Processing techniques to build a Large Multi-Model (LMM) that will mitigate the challenges faced by them & bridging the gap in their educational realm by providing them with highly tailored, context-aware assistance system tailored to work in the college campus setting. This will enable them to interact seamlessly with their peers and other objects within the campus environment. Our model will be powered with facial recognition techniques, deciphering complex scenarios in a localized environment and delivering precise, actionable knowledge about the surroundings to turn daily campus navigation and social encounters into a vivid experience. VisionPulse uplifts the barrier of passive participation in classroom environments by leveraging various techniques to extract audio from lecture recordings and to create them into structured, accessible notes, which will further facilitate dynamic,content-specific Q&A sessions using Retrieval Augmented Generation (RAG). The benefits of this technology ensure that no student feels alienated due to specific barriers in their education by enhancing their class participation through information retention techniques. Besides this, our project aims to provide the user with more features like real-time obstacle detection, spatial location detection,aiding in reading emails, reading, understanding, and conveying text-based instructions in front of the user, enhancing the feasibility and easing the individual's life. VisionPulse will be able to perform actions based on voice commands and be capable of generating audio-based responses as well. Our contribution of VisionPulse to the IIIT Delhi community can act as a gateway in the realm of education for visually impaired students by creating a more inclusive, empowering, and independent educational experience for them. Our project has LLM Project Proposal 2 the potential to become one of the breakthrough contributions in the academic landscape which leverages existing tools and technology to a significant effect, which will help bridge the gaps in education for differently-abled students and create a healthy, inclusive environment where every student can thrive & unlock their full potential to scale great heights & achieve all their dreams.

1 Introduction

Incorporating the emotional dynamics of humans alongside modern methods of computational linguistics and natural language processing is a method which promises significant advancements in how machines understand and interact with humans. Our project addresses this challenge by developing a model capable of pinpointing the exact causes of emotions during conversations. Whether emotions arise from specific statements or are influenced by inherent personal feelings, our model aims to accurately predict these causes and triggers. Our model integrates cutting-

edge NLP techniques like Transformers, attention mechanisms and Convolution Deep Neural networks (CNNs) alongside psychological insights—specifically, the Myers-Briggs Type Indicator (MBTI)—our approach predicts the emotions, emotional causes and triggers based on the conversation. Please refer to background and problem statement sections below for examples of what the model is predicting. Please refer to the problem statement section below for example.

2 Motivation

1. Visually impaired people face a lot of challenges doing usual tasks & are often alienated from activities which most of us can do without much fuss. Hence, ***the primary motivation of our work was to empower the visually impaired people to do everyday tasks at ease & help them overcome the challenges that they face often.***

2. Inform the user whether they can move forward and provide real-time alerts, such as beeping, when obstacles are detected.

3. Since the visually impaired person may not be able to access documents like IIIT Delhi policies, etc., we developed a model that has the knowledge related to IIIT Delhi built within it, which can be accessed and queried by the user at ease.

4. Detect the user's current location and provide guidance on routes to reach their desired destination.

5. Integrate the student's mailbox by developing an email agent integrated with SMTP to enable users to read and write emails using audio prompts eg: "write a formal mail to admin B.Tech inquiring about the last date of fee payment".

6. Allow users to control and navigate the system through voice commands. ***We aim to tailor each feature of this system to enhance its usability on the IIIT Delhi campus.***

3 Problem Statement

Visually impaired students at IIIT Delhi face a significant challenge: the lack of accessible, context-aware educational support tailored to their unique needs. Existing tools often fall short, limiting their ability to seamlessly interact with peers, instructors, and their physical environment. This gap restricts their participation in both academic and social activities.

VisionPulse's Solution:

VisionPulse seeks to address this challenge by harnessing advanced Computer Vision and Natural Language Processing techniques to create a Large Multi-Modal (LMM) system. This innovative solution aims to provide visually impaired students with:

- Real-time obstacle detection
- Location Identification (Visual Question Answering)
- Agent-based system for reading and sending emails
- LLAMA 3 Guard
- General Chat tool
- Student ERP Portal Agent (SQL Agent)
- Agent for retrieving IIIT Delhi Policies

4 Related Work

4.1 Bootstrapping Language Image Pre-training (BLIP)

The BLIP model architecture proposed by Li et al. is a new Vision-Language Pretraining framework that achieves state-of-the-art performance on various vision-language tasks by addressing the limitations of existing methods. BLIP utilizes a new dataset bootstrapping technique called CapFit, which generates synthetic captions and filters out noisy captions to improve the quality of the dataset. The proposed framework introduces a multimodal mixture of encoder-decoder (MED) model architecture and leverages pre-training objectives such as image-text contrastive learning, image-text matching, and image-conditioned language modeling to achieve flexible transfer learning and effective multi-task pre-training. From this paper, we will be using the BLIP model for Visual Question Answering and its various evaluation metrics discussed. The relevant metrics have been mentioned and explained in the evaluation criteria section.

4.2 Pathways Language and Image (PALI)

PaLI (Pathways Language and Image Model) builds on recent advancements in large language models (LLMs) and vision models by jointly modeling visionLLM Project Proposal 4 and language tasks. This approach addresses the growing demand for models that can handle multimodal inputs, where both visual and textual data are processed together. Previous research has demonstrated the effectiveness of scaling models for specific tasks, particularly in language, using architectures like Transformers and Vision Transformers (ViTs) for vision. The model achieves state-of-the-art performance in several tasks, such as image-captioning, visual question-answering, and scene-text understanding. We will be using the PaLI-VQA model for our question-answering tasks.

4.3 LLama 3

The paper introduces a new series of foundational models named Llama 3. These models are a collection of language models designed to inherently handle multilingual tasks, coding, reasoning, and tool utilization. The most advanced model in this series is a dense Transformer with 405 billion parameters and a context window capable of accommodating up to 128,000 tokens. The paper includes a thorough empirical assessment of Llama 3, revealing that it achieves performance comparable to top language models like GPT-4 across a wide range of tasks. Llama 3 is being made available to the public, with both pre-trained and fine-tuned versions of the 405 billion parameter model, as well as the LlamaGuard 3 model, for enhanced input and output safety. Additionally, the paper details experiments where image, video, and speech capabilities were incorporated into Llama 3 using a compositional method, which demonstrates competitive performance with current state-of-the-art techniques in these areas. We will use this model based on various language-related functionalities provided by VisionPulse and will also use different versions of fine-tuned Llama VQA models, such as **MiniCPM-Llama3-V-2_5.**

4.4 Wu-Palmer Similarity Score (WUPS)

The paper by Malinowski et al. introduces a performance measure called the WUPS score for evaluating the quality of system-generated answers. It draws inspiration from the Fuzzy Sets theory and utilizes the Wu-Palmer Similarity(WUPS)score to account for semantic fuzziness between classes. WUPS score penalizes both underestimation and overestimation of answers. The formula considers the intersection of system and ground-truth answers, employing a soft membership measure. Empirical findings suggest a WUP score of approximately 0.9 for precise LLM Project Proposal 5 answers, prompting down-weighting for scores below a threshold. A curve over-thresholds illustrates the trade-off between precision and forgiveness, with WUPSat 0 being the most lenient measure and WUPS at 1.0 equating to standard accuracy. Further details about the evaluation metric will be discussed in the subsequent sections.

5 Real-time Obstacle Detection

The objective is to develop a model that helps any visually impaired individual detect any obstacle in his/her path while walking, so our model will take an image in the input and classify it as **obstacle** or **Non-obstacle**.

5.1 Dataset Collection & Description

Manually collected various images of different locations of IIIT Delhi focusing on obstacles visually impaired individuals may encounter at different body levels.

Head-Level Perspective Images captured at head height to simulate the natural viewpoint, focusing on high-standing obstacles like door frames, signs, branches, etc.

Knee-Level Perspective Images captured at knee height to identify low-lying obstacles such as chairs, steps, small animals, etc.

- Training Dataset: 167 images of IIITD
- Class 0 (Obstacle): 80
- Class 1 (No obstacle): 87
- Testing Dataset: 40 images of IIITD

5.2 Model Architecture and Approach

Teachable Machine: CNN-based model

- Input (224x224x3) $->$ 3 CNN layers $->$ Output (2 classes)
- The preprocessing steps involve resizing the image to 224x224 dimension and normalizing the pixel values between 0 and 1 after dividing by 255.
- It is compiled with the Adam optimizer, binary cross-entropy loss, and accuracy as the metric.
- This is an attempt to use a very rudimentary CNN-based model to give a sanity check that this problem requires an acute understanding of the object in the model.

5.3 Results

The below table shows the results of the training and testing set:

Metric	Training Set	Testing Set
Accuracy	0.9401	0.8750
Total Inference Time	0.6435	0.1614
Avg Inference Time	0.0039	0.0040
False Negative Rate	0.0805	0.0000
Recall	0.9195	1.0000

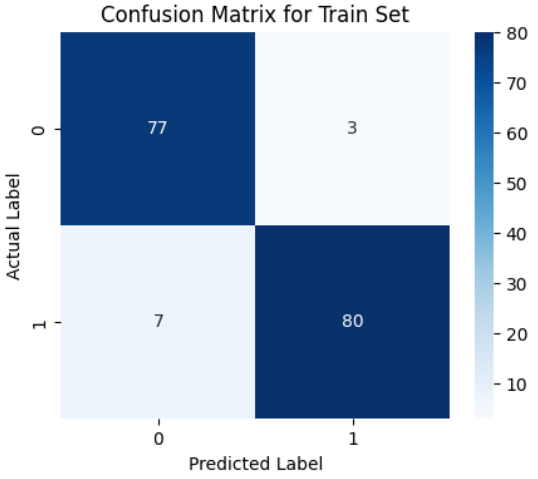


Figure 1: Confusion Matrix (Training)

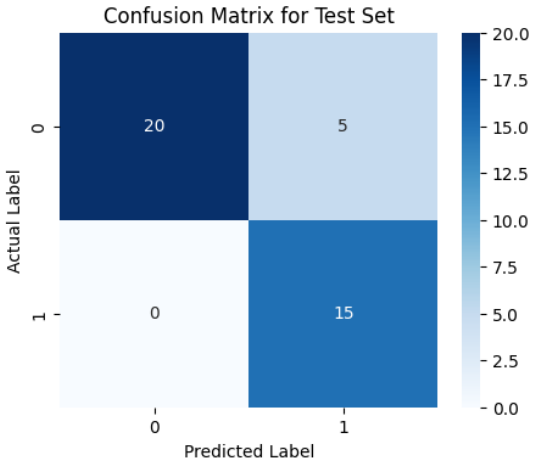


Figure 2: Confusion Matrix (Testing)

6 Receive Emails Tool

To enhance communication and accessibility for visually impaired individuals, we integrate an IMAP server connected to their Gmail account. The system will notify the user of new emails, read the latest email aloud, and convert its content into speech for easy consumption by visually impaired students.

- **Real-Time Email Notifications:** Users are notified instantly when new emails arrive, ensuring they remain up to date without needing to manually check their inbox.
- **Email-to-Speech Conversion:** The latest email is read aloud using text-to-speech, allowing visually impaired users to easily access email content without visual input.
- **Query-Based Interaction:** The system is designed to process user queries related to email content. The Llama 3.1 orchestrator classifies which tool or functionality to use based on the query, ensuring that the right action is taken (e.g., reading specific sections of emails or answering specific questions based on the email content).

Input:

Llama orchestrator: Selects Appropriate tool according to the input query from all available tools.
Once an email is received, it is read aloud, pro-

viding users with easy access to email content without needing to see the screen.

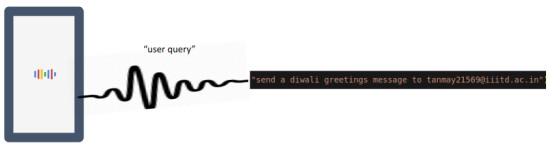
7 Send Emails Tool

The Send Emails Tool integrates an LLAMA model, offering AI-powered functionality to process and send emails efficiently. This tool is designed to streamline email-related queries and automate the process of crafting and sending emails.

- **LLAMA Model Integration:** The tool employs a LLAMA model for interpreting user queries and classifies them to determine if they are email-related.
- **Query Classification:** The LLAMA Orchestrator classifies the nature of the query (e.g., ”send email”) and invokes the appropriate tool accordingly.
- **Message Generation:** The tool accesses the user's mailbox using SMTP and sends the email to the recipient with the help of the LLAMA model. The LLAMA model incorporated at this step, can both send the exact content written by the sender to the recipient or able to generate a message using a query from the sender (For eg: “write an in-depth Diwali message to abc@gmail.com”.

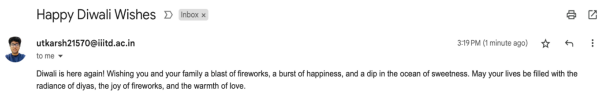
The query from the sender can be made using voice commands, which will be captured, processed directed to the LLAMA model for target generation.

Input:



Llama Guard:

The email is sent to the specified recipient through the integrated SMTP service, as confirmed by the tool.



8 General Chat Tool Functionality

The **General Chat Tool**, implemented as one of the tools invoked by the LangChain agent, incorporates a **LLAMA model** in its implementation.

Key Features:

1. **Fact-based Responses:** The LLAMA model is adapted to provide generic, fact-based responses to the user based on the queries fed into the system. It mimics the functionality of a general chat-based mechanism that interacts with the user and provides responses based on any kind of prompt or query.
2. **Adaptable Query Handling:** The tool allows the user to:
   - Request *basic overviews* of the system.
   - Explore functionalities of the **VisionPulse model**, such as *Real-Time Object Detection*.
   - Ask for a wide range of IIIT-Delhi-based queries or other diverse topics of user interest.
3. **System Integrity Backup:** It acts as a fallback mechanism to maintain the integrity of the system. If a specific query aimed at performing a specific task fails, either due to issues in invoking a tool or retrieving a response, the General Chat Tool provides a generic response to the user. This backup mechanism activates after the system reaches the threshold for

the maximum number of retry attempts allowed (**15 retries, to be precise**).

Thus, the General Chat Tool ensures robustness between UI interactions between the system and the user, by providing adaptable responses and serving as a fallback mechanism to maintain system reliability in case of any potential failures.

9 Student ERP Portal Agent Functionality

The **Student ERP Portal Agent** is highly useful as it has access to the entire ERP system of the user and can provide a variety of responses based on user queries.

Key Features:

1. **Agent within an Agent Framework:** The Student ERP Portal Agent contains an **SQL Agent** within it, designed to provide responses to **Natural Language Queries** provided by the user.
2. **Query Processing and Retrieval:**
   - The query from the user is processed and sent to the SQL agent.
   - The SQL agent formats it into a syntactically correct and coherent SQL query for retrieving the specific information requested, by searching that in the ERP database.
3. **Response Formatting:** The retrieved response, originally structured in SQL format, is:
   - Parsed into a suited format.
   - Propagated back to the main agent.
   - Depicted in the desired format to ensure user readability.
4. **Example Use Cases:**
   (a) "Tell me my grade in 1st semester."
   (b) "Tell me my grade in NLP course."
   (c) "How many credits have I completed till now?"
   (d) "What were the courses where I got a perfect 10?"
   (e) "What are my 2 worst grades and in which course?"

Visualization: Below is an example query from the user to the Student ERP Portal Agent, where it is invoking the SQL Agent to generate responses based on the query:

Thus, the Student ERP Portal Agent improves the user experience by allowing natural language interaction with the ERP system, thereby, ensuring efficient and accurate retrieval of information from the ERP portal. This use case is particularly useful for Visually Impaired Students as they can ask queries using voice-commands and retrieve information in the same format, which helps them to keep tab of important information related to their academics.

10 Python Agent Tool

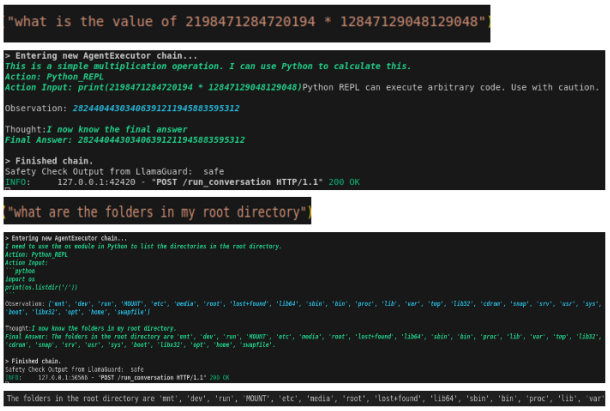


Figure 3: Python Agent Tool

The Python agent tool, referred to as 'python_tool,' integrates VisionPulse access to a wide range of essential and advanced Python libraries. It can perform diverse operations, working on natural language queries provided by the user. This tool can perform tasks like listing files within a directory, generating text files based on user prompts (e.g., creating a text file to track deadlines for a course), and executing complex Python

scripts, including arithmetic operations like long-digit multiplications or additions. This tool can be connected both to the laptop and to the server.

It's applications are extensive. It can be used to generate text files or PDFs, execute system-level updates, solve mathematical problems, etc. It eliminates the need to visually navigate and manage files and other tasks, it bridges the gap in accessibility.