islington college
(इस्लिङ्टन कलेज)

Module Code & Module Title

CS6P05NI Final Year Project

Assessment Weightage & Type

5% FYP Proposal

Semester

2024 Autumn

PROJECT TITLE: CONTEXTUAL OBJECT DETECTION

Student Name: Bibek Poudel

London Met ID: 22067316

College ID: NP01AI4A220032

Internal Supervisor: Alish KC

External Supervisor: Samrat Thapa

Assignment Due Date: Dec 5, 2024

Assignment Submission Date: Dec 5, 2024

Word Count (Where Required):2400

I confirm that I understand my coursework needs to be submitted online via Google Classroom under the relevant module page before the deadline for my assignment to be accepted and marked. I am fully aware that late submissions will be treated as non-submission and a mark of zero will be awarded.

Table of Contents

1. INTRODUCTION TO PROJECT DOMAIN 5
1.1. Problem Statement 6
1.2. Project as a Solution 7
2. AIM AND OBJECTIVES 8
2.1. Aim 8
2.2. Objectives 8
3. EXPECTED OUTCOMES AND DELIVERABLES 9
3.1. Features 9
4. PROJECT RISKS, THREAT AND CONTINGENCY PLAN 10
6. RESOURCE REQUIREMENT 12
7. METHODOLOGY FOR PROJECT DEVELOPMENT 14
7.1. Considered Methodology 14
7.2. SELECTED METHODOLOGY 15
8. GANTT CHART 19
9. WORK BREAKDOWN STRUCTURE 20
10. MILESTONES 21
11. CONCLUSION 22
12. References 23

Table of Tables

Table 1: Risks, Threats & Contingency Plans.....11
Table 2: Hardware Requirements..... 12
Table 3: Software Requirement 13

Table of Figures

Figure 1:Volume of data produced in last decade 5
Figure 2:RUP methodology 14
Figure 3: Prototype Model 15
Figure 4: Agile Methodology 16
Figure 5: Scrum Methodology 17
Figure 6: Gantt Chart 19
Figure 7: Work Breakdown Structure 20
Figure 8: Milestones 21

1. INTRODUCTION TO PROJECT DOMAIN

Artificial Intelligence and technological advancement with the worldwide usage of internet have contributed to the generation of large volume of data all over the web. Social media platforms, smart phones and computer, websites are some of the main reasons behind the exponential growth of the volume of data over the years. According to the report on the worldwide internet usage by (DataReportal, 2024 Jan 31), around 5.35 billion people are actively using the internet in 2024, which contributes to 67% of the worldwide population. This rigorous usage of the internet has generated the large volume of data every day. This hugely generated data have been utilized with the training of the huge AI models such as GPT, Gemini, Gemma and other state-of-art models on computer vision and natural language processing paradigms. Companies like Google, Meta, IBM, OpenAI, and Microsoft have been actively trying to squeeze the maximum potential out of the available data for rapid AI and technological advancement, pushing its development toward Artificial General Intelligence.

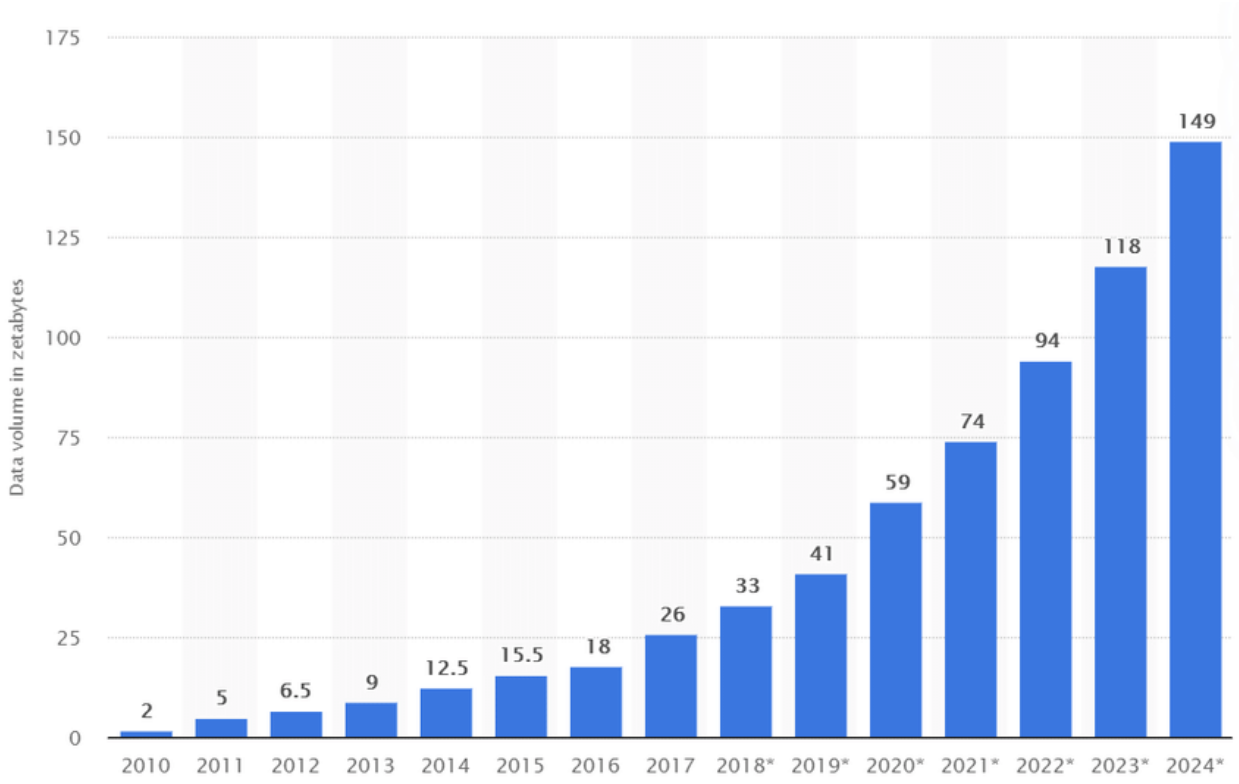

Figure 1:Volume of data produced in last decade

Computer Vision and NLP have their own way of interacting with the data and providing the output for various applications including image classification, image segmentation, autonomous vehicles, unimodal and multi-modal generative models and so on. Both paradigms perform well on their respective areas but the true potential lies between the intersection of NLP and computer vision to allow the models analyze, reason and generate information based on the visually seen data on real time. The leap towards the AGI can be contributed through their integration with minimum complexity in the model architecture and minimum compromise with the performance, computational resources, accuracy and generalization of the model. This project aims to explore this synergy between NLP and computer vision to develop innovative applications while addressing the challenges of model complexity and resource efficiency, ultimately contributing to the broader goals of AI advancement.

1.1. Problem Statement

Object detection has been used as the widely used computer vision tasks such as identifying and localizing the objects from the image or video along with detection in real time scenario as well (Paperswithcode, 2024). Traditional object detection algorithms and state-of-art models generate very little information to be used for sensitive aspects such as providing the traffic state for a visually impaired person to cross the road, information about the intensity of traffic jam and crowd for safe navigation, autonomous vehicles where object detection algorithm often misinterpret the object in the scene causing trouble in real life transportation.

With the world getting complex day by day in forms of infrastructural development and technological advancement, these traditional models found to be underwhelming due to which this project propose the system where we integrate the object detection aspect with the language model so that a reasoning ability is fused inside the classical object detection algorithm. Transportation sector, autonomous vehicles, food manufacturing, biometrics and facial recognition, medical image analysis and video surveillance solutions are the application areas of this project. Although object detection performs well for these application areas, this project aims towards development of the comic-inspired AI powered system that have some (perception, reasoning & speaking) form of human intelligence and ability which is indeed a leap towards Artificial General Intelligence (OpenAI, 2024). This project will showcase the ability to provide the real time context using the visual data as the input along with the ability to read the printed and handwritten text (pre-trained

model) with the application areas towards the assistive device for real time context generation and communication along with an applications area such as navigation and locomotion of robots, robotic arms and autonomous vehicles.

1.2. Project as a Solution

This project proposes a solution to the traditional object detection algorithm with the limitation of context for the detected real time frames by integrating the language-based model with the vision-based object detection model with the finetuning of those models for the tailored output. This integration will also provide the application and research areas in the area of general intelligence in computer vision and natural language processing making possible the architectures like that of as Vision Transformer, Multi-modal models and so on. The developed web application with the model capable of generating the real time context along with extracting the instructions from the image and then communicating based on the provided instructions along with the text-to-speech functionality. This project will solve the challenging issues of traditional object detection model by enhancing and tailoring the integration with language-based model to provide real time reasoning and analyzing capability. Assistive devices for visually impaired individuals will be benefited by this project for safe navigation and perception of what world looks like along with the robots for gaining the intelligent system for surviving in real world environment with numerous action spaces. The large volume of generated data in internet along with pre-trained weights of the open-source models such as LLaMA by Meta, Gemma, YOLO, LLaVA will be utilized for the good cause of accessing the information in the sensitive aspects of the world. Broader sense of intelligence is added along with visual perception and hearing capability will be added to this project solving numerous problems in the area of Artificial Intelligence & Robotics.

2. AIM AND OBJECTIVES

2.1. Aim

Improvement of the existing traditional object detection techniques with the integration of language model for context generation and reasoning. The primary aim of this project is to create a web application with the integration of contextual object detection model along with OCR and voice recognition for robust and efficient performance in real life scenario leading to efficient context generation and usage.

2.2. Objectives

1. Integration and Fine-tuning of existing object detection model with custom dataset for specific use-cases of assistive devices for visually impaired person and guided locomotion and navigation for autonomous robots and robotics arms.
2. Development and integration of OCR and voice recognition with real time context generation and object detection with efficient pipeline for robust performance of the web application.
3. Research, design and development of custom architecture for the contextual object detection model from the areas such as Transformer, visual transformer along with optimizing it for smooth and efficient performance on local devices and edge devices as well.
4. Development of intuitive and expressive user interface and user experience for accessing results of contextual object detection model in a go.
5. Integration of developed and trained model with web application for allowing users to access the actual potential of this paradigm for real world use cases.
6. Collection and preprocessing of the custom dataset required for training the existing state-of-art models for object detection and training the transformer architecture with the custom dataset.
7. Open-sourcing the contextual object detection project for further development and improvement of the project through community support.
8. Intensive and rigorous testing of the trained models in real world scenario to validate the performance of the model in actual use cases.

3. EXPECTED OUTCOMES AND DELIVERABLES

Completion of this research and application oriented final year project will provide numerous useful outcomes including a novel approach to develop an integrated language-based computer vision model, trained contextual object detection model with robust and efficient performance, a web application bridging the gap between the end users and the developed project. This project will be more focused towards reasoning based on the captured frames at real time, the trained model will be able to reason, analyze and answer various queries mainly focused towards the context of the provided scene to the developed system.

3.1. Features

1. Real-time context generation for the captured frames along with the extended capabilities to reason, analyze and answer various queries related to the provided frames.
2. The web app will be integrated with the database containing the information about the various contexts of various object detected scenes for training and validation purpose.
3. Text-to-speech functionality along with voice recognition for communication with trained model about the various captured scenes and frames.
4. Users will also be able to upload the images, videos and generate the context or reasoning about various aspect of frames for human like interaction.
5. OCR functionality for text extraction in real time for digital and handwritten text up to some extent.
6. Assistive device with attention to detail for dealing with the sensitive aspect of the life by visually impaired individuals.
7. An efficient and robust instructions for the navigation and locomotion of the autonomous vehicles with the further enhancement of the project with the edge-compatible model.
8. Intuitive and Expressive UI/UX for interactive usage of the contextual object detection and communication by the end users.

4. PROJECT RISKS, THREAT AND CONTINGENCY PLAN

S. N	**RISKS & THREATS**	**CONTINGENCY PLAN**
1	Resource constraints for handling the large volume of data locally, prototyping with the existing large language models, integration of two different paradigms of artificial intelligence along with real time video processing would require significant computational power from the computer causing the performance bottlenecks along with problems regarding the model loading, dataset processing and integration process.	1. This project will utilize cloud storage and cloud computing for handling the datasets, training the models along with their integration and optimization of the algorithms will be used for system development and prototyping
2	Inference latency during the usage of system and the model with the real-world data, as the higher latency leads to the bad use experience along with improper context generation.	2. This system will utilize the model optimization techniques such as model pruning & quantization to reduce the model size along with inference latency along with proper exception handling for real-world scenario.
3	Issues related to model generalization in numerous real world use cases causing the risks of incorrect context generation and object detection leading to system failure.	3. This system will prioritize the collection and processing of the high-quality dataset with diverse samples to avoid the problem of overfitting and introduce generalization to the trained model.
4	Development of the custom architecture with the vision and language capability would be quite complicated task and could be a serious risk for the project completions.	4. Intensive research will be done with various research papers, forums and portals for the successful development of custom architecture merging two different AI paradigms.

5	Issue with Voice recognition and OCR in the busy scenario along with the performance bottleneck due to the web framework integration.	5. OCR and voice recognition will be tuned with intensive testing on the various use-cases in real world along with optimizing the backend code base for efficient web-app integration.

Table 1: Risks, Threats & Contingency Plans

6. RESOURCE REQUIREMENT

The hardware and software requirements estimated for this final year project is listed below in two different tables.

Hardware Requirements

HARDWARE REQUIREMENTS
CUDA supported computer with Nvidia GPU
16GB of RAM
100 GB of Storage
8GB OF Video RAM
RYZEN/INTEL 12 GEN CPU
USB web camera
Microphone

Table 2: Hardware Requirements

Software Requirement

SOFTWARE REQUIREMENT
Operating System: Linux/Windows
IDE: PyCharm Community Edition
Stable Internet Connection
Programming Language: Python
Frameworks & Libraries: Anaconda, Tensorflow/PyTorch, Django, Cloud Computing Platform (Google Collab), Docker, HTML & CSS.
Gantt Chart: Gantt Project
Research Papers & References: Arxiv.org
Version Control: GitHub
Database: MySQL & Vector Database like Chroma & Pinecone
WBS & Flowchart: Draw.io & Canva

Table 3: Software Requirement

7. METHODOLOGY FOR PROJECT DEVELOPMENT

7.1. Considered Methodology

1. RUP Methodology

RUP is a software engineering and development process that emphasizes software design and development utilizing the UML. Because it enables us to carry out business analysis, design, testing, and implementation software development processes, it aids in the creation of bespoke products. It might assist us in honing our computer programming and project management abilities, which can enhance the readability and adaptability of the code (Indeed Editorial Team, August 16, 2024).

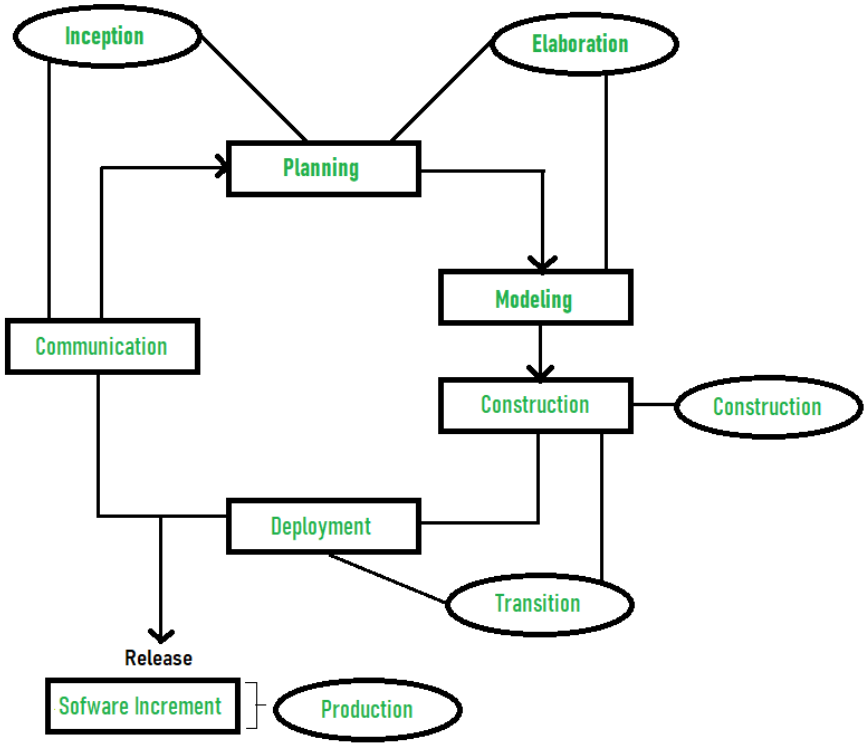

Figure 2:RUP methodology

2. Prototype Methodology

The prototype model is a process of system development where a prototype is constructed, tested, and then modified as needed to provide a satisfactory result. From there, the final system or product is developed. The primary purpose of this tool is to determine whether our project is on the correct course and to maintain user alignment in order to successfully accomplish its destination (Lewis, 2024).

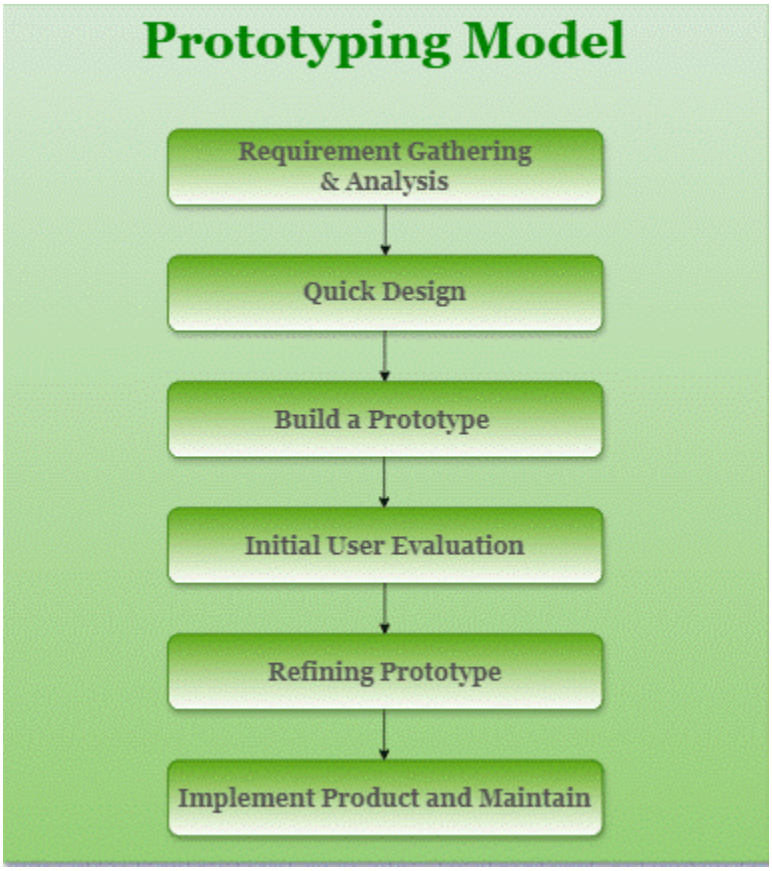

Figure 3: Prototype Model

7.2. SELECTED METHODOLOGY

1. Agile Methodology

For project management, the agile methodology—which divides work into multiple flexible stages called sprints—was chosen from among the solutions that were taken into consideration. It is a process that aims to create, deliver, and test high-quality software as quickly and cheaply as feasible. From a variety of frameworks, I have applied the scrum process here.

Figure 4: Agile Methodology

2. Scrum

The most widely used agile methodology is Scrum. Teams use this management paradigm to self-organize and work toward a common goal. It lists a number of meetings, tools, and duties necessary for a project's successful completion. Scrum approaches can help teams set goals, learn from failures, and adapt to change. It is a well-liked option since it provides a durable and affordable answer to complex problems (AWS , 2024).

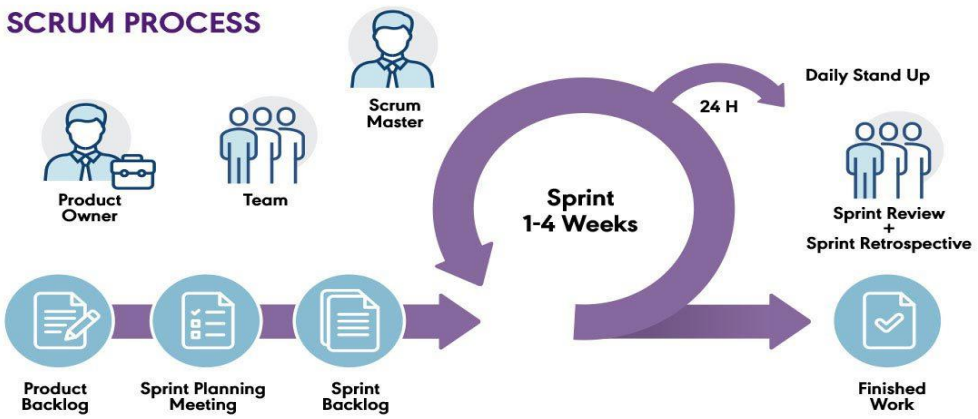

Figure 5: Scrum Methodology

To complete my Contextual Object detection project, I have decided to use the scrum framework. The following are some explanations for using this methodology:

a. **Project Feasibility:** Teams can adapt to changing priorities and requirements thanks to Scrum's iterative methodology.

b. **Visibility of the Project:** Scrum projects are divided into discrete sprints. We can break up our one work into multiple sprints to make it easier to finish our task correctly.

c. **Time management using sprints:** A timed-boxed sprint allows us to focus on a single goal and, at the end of each sprint, assess our progress and modify our plans as needed.

d. **Clearly Defined Milestone:** By using this methodology, we are able to precisely define milestones, which helps us see and monitor our development process throughout time.

8. GANTT CHART

Figure 6: Gantt Chart

9. WORK BREAKDOWN STRUCTURE

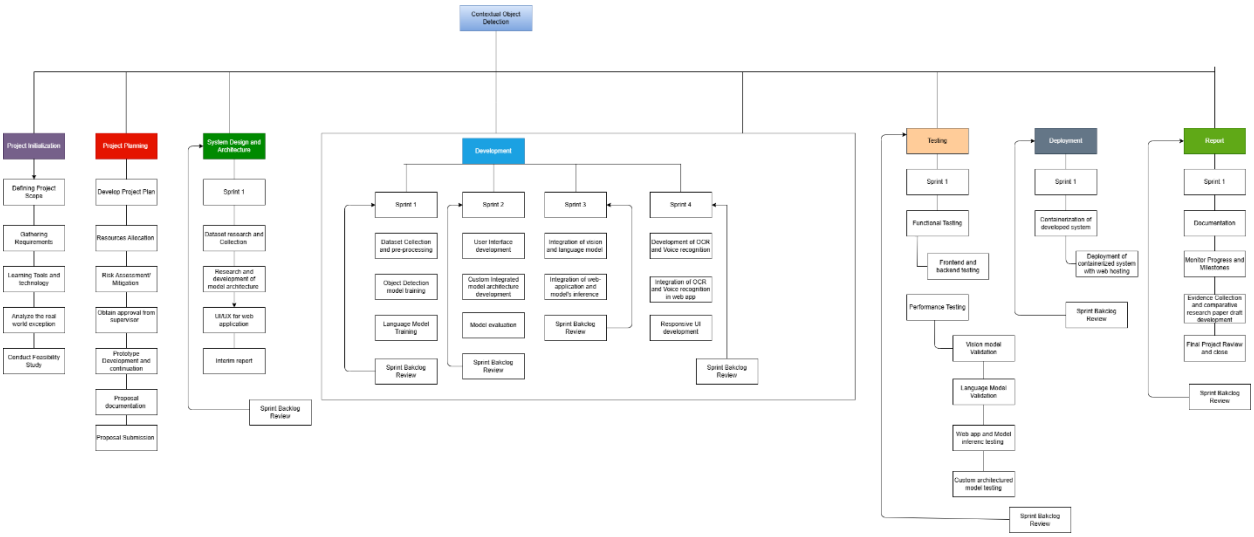

Figure 7: Work Breakdown Structure

10. MILESTONES

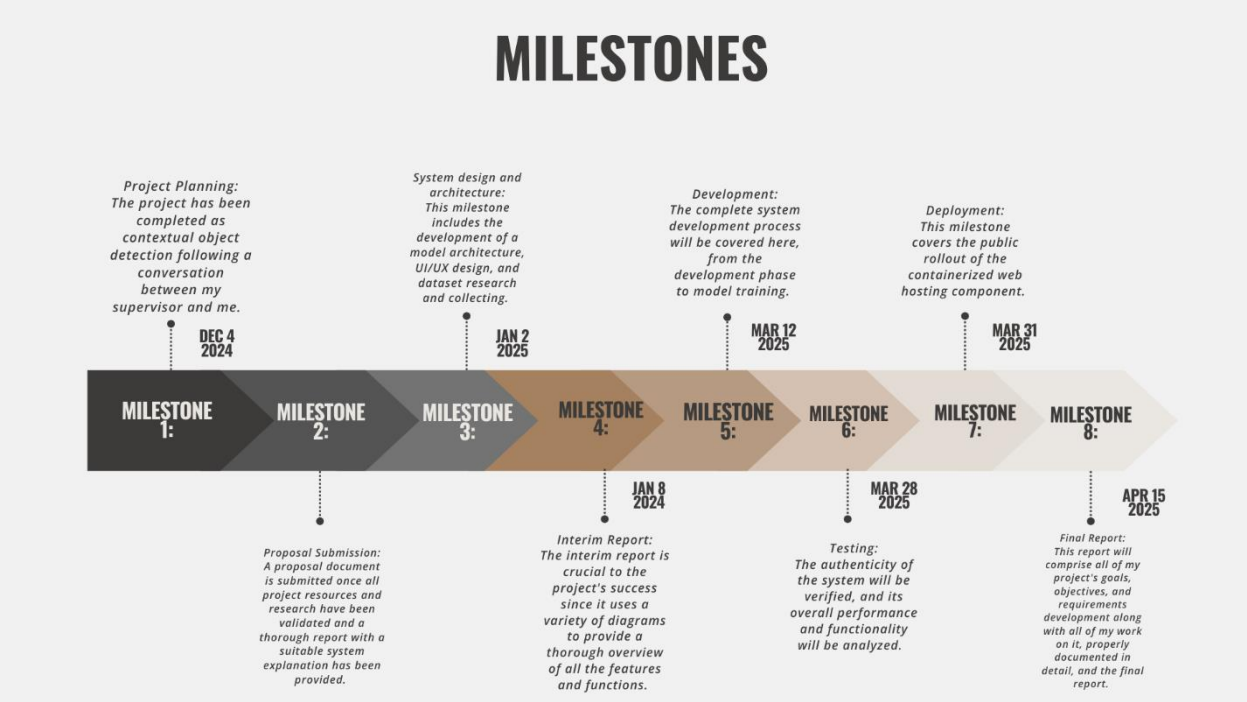

Figure 8: Milestones

11. CONCLUSION

To sum up, Contextual Object Detection, represents an important step towards advancing traditional object detection with the integration of natural language processing for contextual reasoning. The integration of vision and language models in real-time enables not only object detection but also the generation of meaningful, context-aware insights that address critical real-world applications. From assisting visually impaired peoples with navigation and knowing what happening around them along with enhancing the autonomous systems and vehicles filling the gap between human-like intelligence and reasoning with machine perception. With the extensive research and testing on various AI models, tons of research papers, innovative methodologies and robust implementation strategy, the project aims to deliver the novel architecture of real-time context generation from visual data with the additional functionality of OCR and voice recognition for two-way communication between the system and the end users. The proposed system also highlights practical applications and research areas in assistive technologies like Meta glasses, robotics and sensitive aspect of world requiring precise understanding and decision making.

The successful completion of this project, along with the implementation of a robust testing framework and ongoing enhancements driven by community collaboration, establishes it as a significant advancement in the fields of computer vision and artificial intelligence. It highlights the power of collaborative efforts in AI to drive innovation and move us closer to realizing systems that embody the principles of Artificial General Intelligence opening the possibilities for lot of research.

12. References

AWS , 2024. *What is Scrum?.* [Online] Available at: https://aws.amazon.com/what-is/scrum/ [Accessed 4 Dec 2024].

DataReportal, 2024 Jan 31. *Internet use in 2024.* [Online] Available at: https://datareportal.com/reports/digital-2024-deep-dive-the-state-of-internet-adoption [Accessed 04 December 2024].

Indeed Editorial Team, August 16, 2024. *RUP: Definition, Phases, Advantages and Best Practices.* [Online] Available at: https://www.indeed.com/career-advice/career-development/rup [Accessed 4 Dec 2024].

Lewis, S., 2024. *Prototyping Model.* [Online] Available at: https://www.techtarget.com/searchcio/definition/Prototyping-Model [Accessed 4 Dec 2024].

OpenAI, 2024. *Pioneering research on the path to AGI.* [Online] Available at: https://openai.com/research/ [Accessed 4 Dec 2024].

Paperswithcode, 2024. *Object Detection.* [Online] Available at: https://paperswithcode.com/task/object-detection [Accessed 4 Dec 2024].