Developing an Intelligent Virtual Assistant for Extended Reality Environments: A Multi-Modal Approach for Understanding User Intent and Context using Computer Vision, Natural Language Processing, and Large Language Model.

by

Nafis Rayan

21301559

Srejon Ahamed Joy

21301354

MD. Ehtashamul Islam Khan

21101087

Abtahi Shipar

21301233

A thesis submitted to the Department of Computer Science and Engineering in partial fulfillment of the requirements for the degree of

B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering

Brac University

June 2025

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Nafis Ryan	Srejon Joy
21301559	21301559
Ehtashamul Khan	Abtahi Shipar
21101087	21301233

Approval

The thesis "Developing an Intelligent Virtual Assistant for Extended Reality Environments: A Multi-Modal Approach for Understanding User Intent and Context using Computer Vision, Natural Language Processing, and Large Language Model." submitted by

1. Nafis Rayan (21301559)
2. Srejon Ahamed Joy (21301354)
3. MD. Ehtashamul Islam Khan (21101087)
4. Abtahi Shipar (21301233)

Of Spring, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 20, 2025.

Examining Committee:

Supervisor:
(Member)

Dr. Jannatun Noor Mukta

Supervisor

Department of Computer Science and Engineering

School of Data Science

Brac University

Program Coordinator:

(Member)

Dr. Md. Golam Rabiul Alam

Professor

Department of Computer Science and Engineering

School of Data Science

Brac University

Head of Department:

(Chair)

Dr. Sadia Hamid Kazi

Associate Professor

Department of Computer Science and Engineering

School of Data Science

Brac University

Abstract

Virtual reality (VR) and Mixed reality (MR) technologies have the capacity to enhance and revolutionize a variety of industry sectors, providing individuals with greater immersive and engaging interactions. Nevertheless, modern virtual reality technologies lack a virtual intellectual assistant that recognises and responds to users' visual and auditory inputs. Thus hindering users from executing tasks, accessing information, and interacting with the overall virtual system environment. The goal of this research is to develop an immersive and intelligent virtual reality assistant to work with individual users' contextual intent. The proposed Virtual Reality Intelligent Assistant (VRIA) will be able to understand and respond to users' auditory and visual inputs and help users' daily endeavours involving task automation, information retrieval, navigation, object recognition, entertainment, personalisation, education, home automation and communication to enhance overall productivity. VRIA offers the ability to leverage natural language processing, computer vision and large language models for understanding visual and auditory inputs. The virtual ecosystem enables user-friendly, tailored, personalised and highly adaptable VR environments to accomplish real-world endeavours. In addition with the incorporation of powerful UI, IoT devices, Mixed Reality attributes, and Cloud Computing infrastructure, the Virtual Reality Intelligent Assistant (VRIA) will meet the requirement for real virtual world interactions and be scalable and flexible across multiple VR/MR platforms and applications.

Keywords:

Virtual Reality, Mixed Reality, Interactive Experience, Virtual Assistant, Visual Inputs, Auditory Inputs, Task Automation, Navigation, Object Recognition, Natural Language Processing, Computer Vision, Large Language Model, Human-Computer Interaction, Internet of Things, Cloud Computing.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years, Virtual Reality (VR) and Mixed Reality (MR) technologies have revolutionized rapid advancements, transforming the way people interact with virtual content. These technologies have the potential to revolutionize several industries, including gaming, marketing, healthcare, education, and entertainment, by providing consumers with immersive and engaging experiences. However, despite their great capabilities, existing virtual reality technologies often do not provide seamless interaction between users and the virtual environment. This gap is significantly felt in the lack of intelligent virtual assistants that can understand and respond to users' contextual, visual and auditory inputs.

The Virtual Reality Intelligent Assistant (VRIA) concept aims to resolve these problems through the application of modern technology features such as Large Language Models (LLM), Computer Vision (CV), and Natural Language Processing (NLP). VRIA is a Virtual Reality Internet Application. It gives the user a direct and natural way to act, get information, and communicate in a virtual world through the Web. The integration of these technologies makes it easy for users to engage with the system via text, voice, or gestures, and VRIA will dynamically interpret and understand users' context to provide seamless experience on any network.

To conclude, this paper proposes a method for creating a Virtual Reality Intelligent

Assistant (VRIA) using virtual reality in all kinds of VR/MR environments, covering individual's (photo-realistic vision, sound or text), object recognition, navigation and task automation and external data exchange. This is achieved by combining cloud computing, mixed reality and IoT to merge the virtual and physical worlds. VRIA will be a multi-modal platform where it will adopt Computer Vision (CV), Natural Language Processing (NLP) and Large Language Models (LLMs) to learn micro details on the user continually and keep changing based on the user demography. The user-friendly engagements and experiences are aimed to be delivered with the principles of Cloud Computing and strong UI design.

1.1 Rationale of the Study

With the rapid growth of Virtual Reality (VR) and Mixed Reality (MR), humans are experiencing virtual worlds in new and innovative manners. However, current systems are still not very intelligent, not very aware, and lack support for multiple inputs, especially in resource-poor nations like Bangladesh. Most VR/MR systems do not support visual and audio inputs, which restricts their applications in education, training, accessibility, and productivity.

The goal of this project is to engineer VRIA, an intelligent and accessible VR/MR system, capable of surmounting such obstacles. VRIA does not use traditional virtual assistants but uses advanced tools like Vision Language Models (VLMs), Visual Question Answering (VQA), and multimodal data like visual, audio, gesture, and context data to provide intelligent, real-time assistance. It is not only to innovate but also to provide inclusion. VRIA is aimed at serving persons with technical capabilities, persons with disabilities, and persons in less-privileged communities. Through the stress laid on immersive technology, personalization, ease of use, and basic components of User Interface (UI), VRIA will change how virtual environments are interacted with.

For the sake of learning, working, and playing, VRIA helps users to accomplish things, retrieve information, and have more meaningful virtual conversations, allowing the digital world to be easier and more enjoyable for everyone.

1.2 Problem Statement

Although virtual and mixed reality (VR/MR) have made good progress, their systems lack intelligent virtual assistants for more intuitive interaction with users. Because systems do not yet process contextual information in real-time, handling certain multimedia tasks and looking up information is hard for users.

As it is widely noticed, the lack of sufficiently developed advanced features often places virtual-reality (VR) and mixed-reality (MR) devices on the fringe of the daily routine of the users and on the outside of a moderate level of sustained usage. Bringing Computer Vision, Natural Language Processing, Large Language Models, Internet of Things and Cloud Computing into a system that can be controlled and adapted is very difficult. Virtual environments cannot be made fully interactive when advanced skills in object detection, speaking with devices and merging data are missing.

Furthermore, present VR/MR tools seldom help people with disabilities or meet the unique needs of communities in Bangladesh. Because not enough attention is given to local matters, VR/MR technologies are not fully impactful in education, healthcare and public sector. As a result, we need to develop and deploy a smart assistant system that does things quickly, boosts ease of use and supports bigger programs.

Problems	**Solutions and Features**
Limited contextual tasks and information retrieval in Virtual Environments.	Robust intelligent system to understand and respond to visual and auditory inputs. Solve daily problems by enhancing productivity and scalability through seamless task automation and contextual understanding.
Productivity & Entertainment with adaptive, customizable interactions.	Design VRIA as an intuitive, customizable, communicative solution for problem solving, task automation, information retrieval, and adaptability.
Challenges in integrating technological components.	Incorporate powerful User Interface (UI), Internet of Things (IoT), Mixed Reality (MR), and Cloud Computing to provide an adaptive and straightforward engagement experience.
Build a connection between the physical and virtual worlds to create a realistic virtual environment.	Implement cutting-edge Computer Vision technologies with advanced Small/Large Language Models. Integrate real-time object detection, gesture interpretation, and multimodal input fusion to create immersive mixed reality environments.
Navigation challenges within virtual environments.	Intelligent navigation systems leveraging spatial awareness, path finding algorithms, and real-time updates to guide users seamlessly through virtual spaces.
Ensuring robust, scalable and flexible deployment.	Utilize cloud computing and distributed systems with scalability and flexibility to operate across diverse platforms and applications.

1.3 Research Objectives

Our goal for this research is to build VRIA, an intelligent, pleasant and powerful virtual assistant for use in VR/MR. Using Computer Vision, Natural Language Processing and blending Small and Large Language Models, VRIA will help users easily understand everything they see and hear.

The assistant software responds to images, identifies movements and actions and gives recommendations in real-time. The aim is to create a high-performance model that can process vision-language fast and energy-efficiently on small devices. It is important that this research introduces VRIA so it can be used by individuals with disabilities through special interaction methods, voice and body movements and access options. Using information on how users interact, VRIA makes sure the experience is adapted and easy for all users. Since physical and digital environments can use the Internet of Things and cloud services, there will be no problems between them.

Designing the information system according to local needs is an essential aim of this study in Bangladesh.It also aims at revolutionizing areas like learning, healthcare, gaming, and business using immersive 3D technology as a component of the assistant system. The book will consider how VRIA can be put to use for building simulations, making prototypes, and testing new ideas in several industries. The use of prototypes in real-life simulations will be used as an example to promote the use of VR and AR in terms of engagement, utility, and accessibility.

1.4 Methodology in brief

VRIA uses a complex, multi-method framework with Computer Vision (CV), Natural Language Processing (NLP) and Large Language Models (LLMs) to offer rich features in VR/MR settings. The VRIA will use text, speech and gestures, combined with multi-modal system to interpret the things people say or do.

The YOLO [58] and TensorFlow [5] technologies are used in the CV module to detect objects and gestures as they occur and NLP is applied to manage by words/text input by recognizing user intents, finding important entities and handling dialogue. LLM integration helps produce responses that depend on the user's context and allows information to move easily between different forms of communication. All the analysis done by each module is fused and this formation is given to the decision-making engine to guide a suitable reaction.

Flask, Three.js and other tools used to develop environments that are engaging and user-friendly according to UI standards. VRIA is supported by nearly all VR/MR headsets such as Oculus, HTC Vive and Microsoft HoloLens, so it has a wide user base. In addition, the simulation system will be made with scalable cloud technology to support a great number of users and future updates, making sure it stays strong, adapts, and is both realistic and engaging for everyone.

1.5 Contribution

A newly designed Virtual Reality Intelligent Assistant (VRIA) is proposed in this thesis, improving both the practical-use and entertainment value of virtual environments. The main achievement is enabling all major technologies like AI, CV, NLP, IoT, MR and Cloud Computing to work together, so the system answers to visual, textual and spoken language inputs at the same time. The VRIA helps fix the drawbacks of VR and MR technologies by making it possible to complete tasks in the right situation, access user data and respond to users using different sensors and systems.

It addresses both productivity and entertainment problems on all platforms by giving users experiences that are customized, simple to use and sharing options that suit their qualities and habits. The system meets the challenge of uniting a variety of technology by organizing them into separate modules and using technology that

allows for flexibility and operation on multiple platforms and uses. Because it can detect objects in real-time, identify gestures and understand what you say, the assistant helps you interact with virtual content more naturally. With intelligent spatial navigation, users are guided efficiently through virtual areas using algorithms and by keeping track of what's happening in their environment. Because of these contributions, tasks are performed better, interaction is more natural, users are more engaged and AI, Extended Reality (XR) and Human-Computer Interaction all move forward. The research prepares for future VR/MR platforms that are both intelligent, accessible and designed for easy user interaction with virtual spaces.

Chapter 2

Literature Review

2.1 Overview

This literature review talks about the integration of Large Language Models (LLMs) and Extended Reality (XR) technologies for their application in immersive human-computer interaction. It also determines how XR (VR, AR, MR) can build on the reasoning, conversational, and multimodal capabilities of LLMs in an effort to enhance virtual assistants, educational software, and user interfaces. The review touches on existing challenges and opportunities in their unification, particularly voice-based and context-aware interactions.

2.1.1 Introduction to Intelligent Extended Reality (XR)

The recent years have seen a rapid growth in the field of Artificial Intelligence (AI), Extended Reality (XR) systems, Virtual Reality (VR), Augmented Reality (AR) and Mixed Reality (MR). The integration of large language models (LLMs), computer vision (CV), natural language processing (NLP) and multimodal sensing is remodelling the way in which immersive environments enable user interaction, learning, simulation, and automation. Projects like VOXReality [32], the primary application platform for Eloquence [1], and virtual agent education systems [49] point out the growing need for XR systems not only to respond to user actions but

also to dynamically identify context, intent, and environment.

Wang et al. [21] research how AI contributes to improving interaction in VR games with improved tracking and haptic technology.Understanding artificial intelligence to be one of many key technolo- gies to enhance the gaming experience, intelligent gaming solutions are increasingly required in the gaming industry. They investigate the utilization of artificial intelligence to strengthen tracking and force/tactile interaction systems in VR games, with emphasis on both software and hardware improvements. It also presents the concept of human-machine hybrid enhanced intelligence, which combines human cogni- tion models to increase traffic perception and decision making in virtual reality games, making interaction more real and friendly.

2.1.2 Multimodal Interaction in Immersive Environments

The VOXReality project [32] is an analytical undertaking closely aligned with the VRIA system objectives. It was designed with funding from the European Commission .It employs NLP and CV technologies jointly to build multimodal XR environments with enhanced human-machine interaction. Nevertheless, the project does not provide adequate data about technical structures and model performance metrics.Moreover this is just a proposed concept, not practically implemented.

A multisensory rendering system is essential to providing user immersion. ScienceSpace and Infed employ visual, touch, and hearing routes to teach up-to-date physical laws. Lykke's work [8] discusses the importance of physical properties, like mass and inertia, for greater spatial presence in VR, citing the challenge of one-to-one correspondence between actual and virtual objects. Haptic retargeting is proposed as an avoidance of physical realism mismatches. Task-oriented multimodal rendering, as described by [1], guarantees that intermodal integration (touch, hearing, sight) enhances comprehension, but most current systems are context-sensitive and rigid.

The Eloquence system (a computational or AI system designed to generate natural,

contextually responsive, and engaging multimodal responses predicts contextual responsiveness to generate more natural and engaging multimodal responses. With continued progress, differences between real-time rendering and accurate psychophysical response model-ling remain a persistent challenge.

Iki and Aizawa [13] examine how an expansion of visual ability in Vision-and-Language (VL) models affects language understanding. They conclude that more expressive visual features can degrade language performance, although VL pretraining reduces it. The study shows the need for well-balanced pretraining to optimize multimodal task effectiveness. Mohamed and Sadeghi [35] propose a collaborative interface that integrates AR and VR based on gesture recognition and multimodal interaction. Their solution enables natural real-time collaboration using head-mounted displays and mobile devices with applications in design and urban planning.

2.1.3 Intelligent Virtual and Augmented Assistants

Research on AR-based virtual assistants complements interaction with multimodal perception. Chen et al. [24] suggest an AR visualization system that is capable of eye tracking and multimodal feedback. Simiscuka et al.[10] create VRITESS, enabling IoT control in real-world VR, a prime example of haptic, audio feedback, and environment synchronisation.

Gesture control continues to be a common interaction technique. Mid-air micro-gestures [15], which were presented as a relief for fatigue induced by handheld controllers, allow for natural and ergonomic interaction. There are also mentions of issues like accuracy of feedback, gesture compatibility, and ergonomics.

Wang [40] underscores AI's contribution to content automation in VR space. Gebczynska-Janowicz [12] advocates for the application of VR in design education, finding students favoring virtual learning settings. Kasyanov et al. [7] illustrate VR/AR's

contribution to learning languages by virtualizing culturally relevant settings to support student understanding and engagement.Tiefenbacher et al. [2] criticize typical AR usability evaluation, proposing a user experience taxonomy and an evaluation framework that considers tracking, rendering, and hand ergonomics.

2.1.4 Educational Applications of LLMs and XR

Education applications show the potential of XR in improving understanding and interaction. RAG-augmented VR learning environments [49] support learners to experience dynamically adapted learning content. Bozkir et al. [43] focus the education potential of XR but caution against hallucination and adversarial vulnerabilities in LLMs. Yousri et al. [60] emphasize MR-based customized training systems using mobile and wearable devices.

Meske et al. [20] emphasize social interaction in VR across industries. Chen and Wang [26] demonstrate the effectiveness of human-computer interfaces using VR in mobile learning. Villani et al.[3] depict MR environments that support multitasking and remote robotic control. The applications of MR provided by [18] reveal ethical and psychological concerns, including data secrecy and reliance.

2.1.5 Virtual Assistants and Human-Robot Interaction

Li et al. [30] suggest "Max," an optimized IVA for industrial robots through REST APIs, natural speech, and prediction models for improved shop-floor conversation. Yadav et al. [41] present a speech-to-text, text-to-speech, and API-integrated-based IVA system but are non-scalable and do not include privacy analysis.

Lu et al. [16] summarize AR-based campus guides apps, noting the immersive benefits of IoT tracking and 3D modeling for outdoor-indoor navigation. Zhang and Smith [11] combine NLP and CV to semantically search video game content to help researchers in games and gamers alike.

Chen et al. [25] propose an instance-level user intent recognition system for iOS apps with accuracy of 64.43% without metadata, even with a limited sample size. Luo et al. [31] introduce Valley, a multimodal foundation model learning from video, text, and image data through a ChatGPT-based pipeline with a two-stage method to efficient video understanding.

AR, VR, and NLP are transforming user experience (UX) evaluation by replacing traditional surveys with immersive feedback mechanisms. Haque and Ameen [42] depict how combining user feedback with algorithmic analytics offers instant feedback, driving product development with more accurate and people-oriented evaluations.

2.1.6 Voice Assistants and Natural Language Interfaces

Voice interaction systems have come a long way from basic voice recognition programs to sophisticated virtual assistants that are capable of communicating with users through natural language. They are increasingly being incorporated into XR environments for more engaging and naturalistic interaction.

Manojkumar et al. [33] provide a systematic review of Python-based AI virtual assistants operating in supervised, unsupervised, and reinforcement learning modes. These systems are capable of performing tasks such as checking the weather, accessing online content, and opening desktop applications. The authors, however, noted critical limitations in data reliability in modules such as Wikipedia, where the assistant can give incomplete or outdated information, undermining overall accuracy and reliability.

In a related effort, Kumar et al. [29] develop a Voice-Based Virtual Assistant (VBVA) that enables natural command execution with voice recognition. The system allows increased user accessibility and interaction, but also poses questions about the scalability and high-performance capability of the assistant in extensive or diversified tasks.

On mobiles, Deshpande et al. [27] suggest the Intelligent Companion App, a virtual helper implemented on Android using Dialogue Flow. It has conversation interaction and conducts day-to-day tasks such as calling, volume, and weather reports. The app provides a more human-like, interactive experience to the users, justifying the significance of personalization in voice interfaces.

These modern systems are a culmination of the legacy of early voice recognition systems. Machines like IBM's Shoebox, which could understand 16 words, and CMU's Harpy, which could understand over 1,000 words, were the forerunners of today's popular voice assistants like Siri, Google Assistant, and Alexa [17]. Intelligent Virtual Environments (IVEs) integrate VR and AI to offer interactive situations where users interact using natural language. NLP allows natural communication, thus resulting in more user interaction. While IVEs [4] hold great promise for education and entertainment, linguistics is overly complex, making it difficult to create reliable NLP systems.

Su [37] introduces Voice2Action, a state-of-the-art voice interaction system using language models that converts voice commands into actions in real time inside 3D environments. It consists of preprocessing, classification, extraction, and execution modules. It also includes a hierarchical processing model for reducing error rates and enabling feedback-based correction, making it highly suited to VR tasks where keyboard or touch input is inconvenient.

2.1.7 Integration of LLMs in Mixed Reality Development

Fernanda et al. [28] introduce LLMR, a Unity-compatible large language model framework to generate interactive 3D scenes in real time. The system employs modular modules such as Scene Analyser GPT to interpret the environment and user inputs and refresh the virtual world dynamically. Bozkir et al. [43] and Yousri et al. [60] propose frameworks integrating LLMs with mixed reality (MR) devices to facilitate adaptive, personalized learning and feedback provision. The systems

utilise LLMs' contextual reasoning and multimodal input processing to enhance user interaction and inclusivity in learning environments.

Konenkov et al. [53] develop VR-GPT, a system that unifies custom visual language models (VLMs) with Unity to enable instruction of tasks in virtual worlds using natural language instructions. This renders interaction more natural and simpler, particularly for complex tasks in virtual reality environments. In the biomedical field, LLaVA-Med, as presented by Li et al. [59], extends the capabilities of GPT-4 to visual question answering (VQA) in medical image tasks. The model demonstrated superior performance compared to baseline approaches, making it a significant advancement in AI-assisted diagnostics.

Buschoff et al. [70] improve reasoning in multimodal LLMs, identifying deficits in physical causality understanding, spatial understanding, and human mental model understanding, essential features in coherent interaction in virtual or mixed reality settings. Carnes and Casado-Mansilla [22], by contrast, compared the performance of cloud-based NLP services AWS, Azure, and IBM Watson for VR applications. Their findings showed that AWS excelled in both part-of-speech (POS) tagging and object description, making it the most excellent of the services examined for processing language in virtual reality.

Rohith et al. [36] indicate the use of Large Language Models (LLMs) like ChatGPT in aiding VR development. Such platforms enhance code quality and ease of use for novice developers but raise issues regarding code reliability. The study emphasises human intervention in LLM-augmented development processes

2.1.8 WebXR and Ergonomic Virtual Prototyping

WebXR enables native support for immersive extended reality (XR) experiences in web browsers, which means cross-platform accessibility without the need for particular applications. Lee et al. [14] critically discussed the WebXR Device API, noting

the ease of deployment to devices and operating systems due to native browser support. Chauhan [23] examined the intersection of XR interface design and ergonomics, emphasizing the necessity of virtual prototyping for demographic-specific UI testing. He also highlights the value of cost-efficient, early-stage prototyping using immersive virtual simulations to accelerate iterative design cycles and enhance usability outcomes.

In addition, Pfeiffer and Pfeiffer-Leßmann [6] offers a novel means of creating mixed reality–Internet of Things (MR-IoT) UIs by enabling physical UI prototyping in XR environments. This reduces hardware-specific development tool dependency and allows for rapid experimentation and optimization of interface components in spatial computing environments.

2.1.9 Object Detection and Visual Intelligence

Real-time detection of Near-Real-Time objects is paramount in enabling responsive and context-aware environments in extended reality (XR) applications. YOLO-based detectors, like YOLOv10 [71], offer double assignment methods and NMS-free models, which further improve inference speed and detection accuracy. Vaishnavi et al. [39] compare YOLO with Faster R-CNN and observed performance trade-offs between the speed of detection and resolution of the scale, for which YOLO excels at real-time responsiveness and Faster R-CNN offers better-resolution object localization.

Prabu et al. [9] conduct an in-depth examination of 21 variants of DEtection TRansformer (DETR) models and identify significant advances in query planning and transformer-attention mechanisms that enhance object detection performance for immersive and dynamic environments. Coupled with these detection models, the ABC-CNN method unites computer vision (CV) and natural language processing (NLP) for visual question answering (VQA) tasks. Through attention mechanisms employed to attend to image regions of interest, ABC-CNN enables multimodal

understanding.

These detection and VQA models are increasingly being utilized in accessibility tools, dynamic interactive systems, and adaptive interfaces in XR to facilitate real-time interpretation of the environment, user assistance, and personalized interaction paradigms.

Lastly, while there has been tremendous progress in the development of intelligent extended reality (XR) systems, there remain several issues of concern. Large language models (LLMs) are still susceptible to hallucinations [43], are subject to security threats, and are non-transparent in their decision-making [70]. Moreover, the scalability of intelligent virtual assistants (IVAs) and virtual personal assistants (VPAs) across various devices and use cases remains limited [25]. Ergonomic fatigue with gesture-based interaction systems also continues to influence user comfort and long-term usability [15]. These constraints can be addressed through the establishment of robust testing architectures [2], enhancement of cross-platform API functionality [14], and the application of open model reporting practices [32]. Future studies must prioritise, with the utmost importance, the development of inclusive, secure, and adaptive XR systems that are highly capable of fulfilling the requirements of diverse users. As immersive computing pervades more of our lives, the future will be shaped by ethically aware design, cognitive user enhancement, and delivering tailored, multimodal experiences that together embody sensibly truly intelligent XR environments.

2.2 Comparative Analysis

We conducted our quantitative analysis by reviewing a wide range of published studies. These papers employed metrics such as system performance, interaction accuracy, task completion rates, and user satisfaction ratings. Moreover, these studies often used controlled user experiments, survey-based evaluations, and statistical comparison techniques to assess technologies like Intelligent Virtual Assistants (IVAs), Augmented Reality (AR) navigation tools, and multimodal systems integrating NLP and computer vision. For example, the "Valley" [31] multimodal model and LLaVA-Med [59], performance was quantified using accuracy scores on specialised datasets benchmarks such as biomedical VQA and video understanding. Evaluation frameworks for tools like WebXR and Voice2Action included comparative metrics like F1 Score and response latency. These numerical data help substantiate claims regarding the efficiency and usability of these systems across varied real-world scenarios.

Research/Study	Private or Public Dataset	VR/AR Implemented	Image Processing	LLM/NLP Implemented	Input Type	Language (Programming)	Applied
Maniati et al. (2023) [32]	Not specified	Yes	Used	LXMERT & UNITER	Text, Video	Not Specified	✗
Konenkov et al. (2024) [53]	Collected Dataset (Private)	Yes with Unity	Used	Qwen-VL	Audio, Text	C#	✓
Yousri et al. (2024) [60]	Not specified	Yes	N/A	PaLM2	Image	Python, FastAPI, PostgreSQL	✓
Izquierdo-Domenech et al. (2024) [51]	Not specified	Yes	N/A	Retrival-Augmented Generation (RAG)	Text	C#	✓
Zhang & Smith (2019) [11]	Gameplay videos on YouTube (Public)	N/A	Yes (Inception)	fastText model (NLP)	Text, Image	Not Mentioned	✓
Iki & Aizawa (2021) [13]	GLUE benchmark (Public)	N/A	Yes (V&L)	N/A	Primarily Text	Not Mentioned	✗
Lu et al. (2021) [16]	Campus navigation dataset (Public)	AR with Unity	Used	N/A	Visual inputs through mobile cameras for navigation	C#	✓
Chen et al. (2023) [25]	Created user interface dataset (Public), Common Mobile App/Web Icon from kaggle (Private)	N/A	Yes (CNN icon classifier)	BERT (64.43%)	Verbal, Textual	Python (for machine learning and data processing)	✓
Kumar et al. (2023) [29]	A new voice dataset was collected	N/A	N/A	GPT Language Model	Audio	Python for implementing voice contorl and processing functionalities	✓
Luo et al. (2023) [31]	Constructed a dataset comprising 73,000 video-based instruction data(Public)	N/A	Used	Stable-Vicuna as the LLM backbone(84.82)	Text & Video Data	Not Mentioned	✓
Manojkumar et al. (2023) [33]	Not specified	N/A	N/A	Yes (But not mentioned)	Audio	Python	✗
Bozkir et al. (2024) [43]	Not Mentioned	Yes	N/A	GPT-4	Text, Audio	Not Specified	✗
VRIA (Our Proposed)	Own Dataset (Private)	Yes with ThreeJS	YOLOv11	Gemma2 9b	Text, Audio, Visual Input	JavaScript, Python, HTML, CSS, Tailwind	✓

Table 2.1: Comparative Overview of Research Studies

2.3 Quantitative Analysis

We conducted our quantitative analysis by reviewing a wide range of published studies.These papers employed metrics such as system performance, interaction accuracy, task completion rates, and user satisfaction ratings. Moreover, these studies often used controlled user experiments, survey-based evaluations, and statistical comparison techniques to assess technologies like Intelligent Virtual Assistants (IVAs), Augmented Reality (AR) navigation tools, and multimodal systems integrating NLP and computer vision.For example, "Valley" [31] multimodal model and LLaVA-Med [59], performance was quantified using accuracy scores on specialised datasets benchmarks such as biomedical VQA and video understanding. Evaluation frameworks for tools like WebXR and Voice2Action included comparative metrics like F1 Score and response latency. These numerical data help substantiate claims regarding the efficiency and usability of these systems across varied real-world scenarios.

2.4 Qualitative Analysis

Qualitative insights were gathered from literature surveys and case studies that documented human experiences with emerging virtual and augmented reality interfaces. Papers like those on immersive learning with LLM-VR fusion or human-machine teamwork in industrial robotics have dense anecdotal evidence and open user feedback. These narratives revealed themes like the need for personalization, intuitiveness of interaction, and emotional resonance of immersive experiences. For thematic analysis, recurring concepts like "engagement," "natural language interaction," and "cognitive load" were extracted from user interviews and case descriptions. These were coded and cross-compared between studies, allowing us to build a detailed image of the human factors issues of concern in these technologies, and ultimately allowing an advanced interpretation of system design impacts on user behavior and perception.

Chapter 3

Working Plan

VRIA is an Intelligent Virtual Assistant for Extended Reality Environments using a multi-modal approach containing Computer Vision (CV), Natural Language Processing (NLP), Large Language Models (LLMs) to capture user intents and context. The system architecture brings together several modules to provide a complete understanding of user intent. It is also built for smooth integration with Virtual Reality and Mixed Reality (VR/MR) platforms, making it easy for users to adopt. This approach seeks to help end users and software companies by encouraging wider use and practical application in different areas.

3.1 Multi-Modal Approach

Enabling a multi-modal paradigm means linking of Computer Vision (CV), Natural Language Processing (NLP) and Large Language Models (LLMs). In this work, we look at two ways to enable a multi-modal interaction paradigm. The first approach uses Visual Language Models such as Phi-3-Vision [57], LLaVA [59], DeepSeek-VL [55], and smolVLM [66], which combine vision and language into one model. These models can perform activities like object recognition, gesture recognition, and visual cue interaction, along with performing Visual Question Answering (VQA), where the model reads out an image and responds to questions about it in natural language. This approach provides a highly integrated and non-disjoint user experience

by enabling the system to understand and respond to inputs with both text and visual content. However, the main drawback of this approach is its high resource utilization, so it would be less preferable to deploy on low-resource systems.

The second approach uses a modular structure where a Computer Vision (CV) and a Large Language Model (LLM) run separately but are interlinked through data exchange. Here, the CV model processes the visual inputs, extracts relevant information for the task to be performed, and sends the context to the LLM, which performs tasks based on language such as Visual Question Answering. While this method will not benefit from the same degree of tight integration as the first approach, it offers a more resource-limited alternative. It still supports multi-modal interaction by letting the system receive visual input and produce natural language output. This makes it a practical solution for situations with limited computational resources.

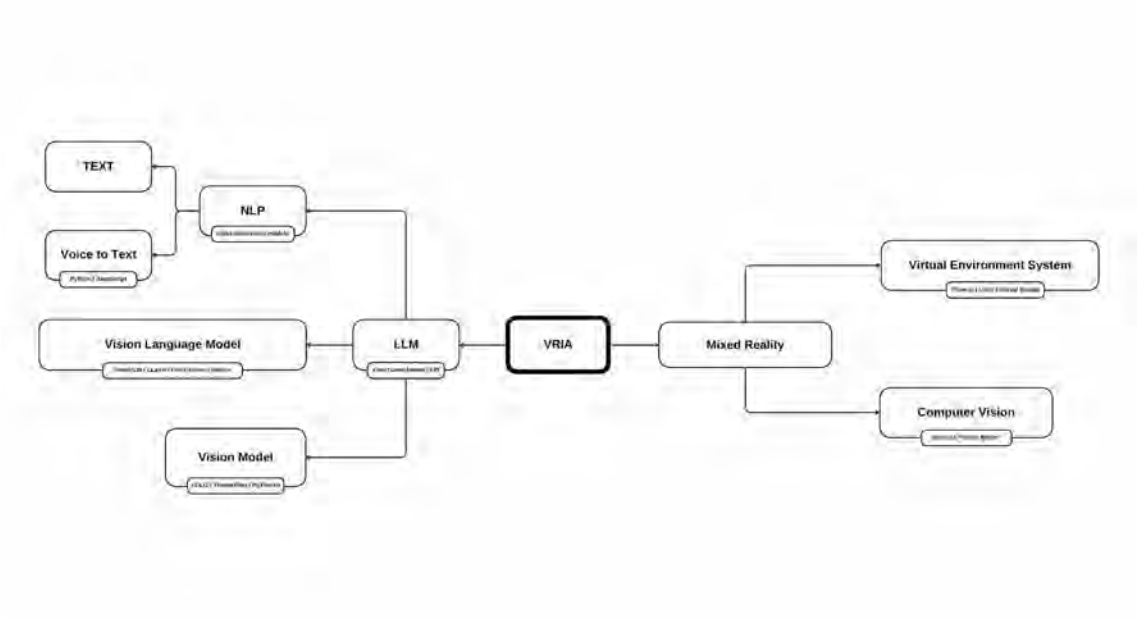

Figure 3.1: Working Plan

3.2 Understanding User Intent and Context

In order to effectively determine the intent and contextual state of the user in extended reality contexts, the VRIA will interpret and process both visual and voice commands. This is to be done via a synergy between planning, execution, inspection, and learning which will ultimately be processed by vision and language model synchronization. By synergizing these methods, VRIA can adjust its interpretations and respond appropriately based on the context at any moment. This helps the assistant stay relevant, responsive, and supportive during all types of interactions. Such a combination of Computer Vision, Natural Language Processing, and Large Language Models enables the VRIA to interpret sophisticated real-world situations and calibrate its responses based on the precise requirements, linguistic subtleties, and input modalities whether acquired by way of speech, text, or visual input.

3.3 Natural Language Processing (NLP) Module

The NLP module is a significantly crucial module in the VRIA since it serves as an interface between the computer vision model and the large language model. The module enables the system to understand and react to voice commands and natural language text input prompts from the user. Using techniques such as intent recognition, entity extraction, and dialogue management, the NLP module renders the user's interactions natural, interpretable, and contextual. The NLP module converts user input in the form of text and speech into structured data consumable in order to align the visual perception of the world with the response generation capability of the LLM. This enables the VRIA to handle commands and queries effectively and intelligently, providing free and uniform interaction over a range of input modalities. A fusion mechanism is used to fuse information from Vision Model and Language Model together in order to understand user intent and context in a coherent manner.

3.4 Computer Vision (CV) Module

As part of the overall mission to expand perception and interaction in Virtual and Mixed Reality experiences, the Computer Vision (CV) module greatly expands the functionality of the Virtual Reality Intelligent Assistant (VRIA). It leverages state-of-the-art technologies such as YOLO (You Only Look Once) [58] for real-time object detection and recognition with the ability to train custom object detection models to accommodate niche application use cases. Even more advanced multimodal comprehension can be achieved through the incorporation of the newest vision language models such as SmolVLM [66], Phi-3-Vision [57], DeepSeek-VL [55], and LLaVA [59]. These enable the VRIA to interpret and react to visual data, objects, gestures, and environmental inputs extremely intuitively and contextually. Furthermore, by utilizing the CV models being trained on custom datasets, the system is able to generate even more accurate and specific responses, depending on the specific application or user case. This flexibility not only works to create the feeling of naturalness of the system but also means users are presented with responses that are not just correct but contextual, leading to the overall effect of being in advanced, immersive environments.

3.5 Large Language Model (LLM) Integration

The VRIA has a robust Large Language Model (LLM) that is instrumental in comprehending user intention and generating smart, context-sensitive responses. The LLM processes not just explicit user input (text, or voice) but also uses fused information provided by the NLP module and the Computer Vision system. Working over this combined input, the LLM makes responses semantically precise, linguistically appropriate, and contextually specific to the user's situation of use. Such an integrated architecture enables natural, cognitive dialogue where the system understands the surface and implicit meaning of user interaction spanning modalities. New technologies like Mistral AI, Deepseek, Microsoft's Phi and Phi-Vision, Meta's

LLaMA, Google's Gemma and Gemini are building similar ecosystems with their work on driving innovation in small, efficient, and ethical language models. These innovations are making it possible to develop high-performance AI systems like VRIA in a cost-effective and scalable manner without the exorbitant cost of computation that large models usually have.

3.6 User Interface and Interaction Design

To provide a smooth and interactive user experience in the VRIA system, careful incorporation of contemporary web technologies and powerful User Interface is necessary. The user interface shall be created with HTML, Tailwind CSS and Three.js to achieve simple, responsive, and visually pleasing layouts with ease and consistency on different devices. For routing and backend, Flask will be utilized along with JavaScript as a lightweight web framework that is extensible and will easily combine the Frontend and AI modules of VRIA. For incorporating immersive VR interactions within the interface for real-time 3D environment and visual element rendering within the browser, Three.js will be utilized. This will allow natural interactions for users, such as an intuitive and dynamic interface. In addition to this, incorporating powerful UI will guarantee that the interface is accessible, usable, and effective irrespective of the input modality utilized whether it is voice, gestures, or typed commands. When it comes to creating the virtual or mixed reality worlds themselves, environments like Unity or Unreal Engine can be leveraged to deliver high-fidelity immersive experiences. Cumulatively, this stack of technologies allows for building a modern, responsive, and novel user interface by which individuals can interact with VRIA in a natural and effective manner.

3.7 Immersive Environment and Scalability

In order to ensure wide reach and ease of compatibility, the VRIA developer kit will be usable through major VR and MR platforms like Oculus, HTC Vive, and Microsoft HoloLens through standardized methods. This enables VRIA to reach out to a large number of platforms and types of applications, such as voice-based, gesture-based, and text-based interfaces. Simultaneously, the creation of immersive environments is a top-level priority. Three.js for 3D rendering within the browser will be used to make visually appealing and interactive environments. The selection of an appropriate engine based on performance, ease of use, and VR hardware support will render the environments interactive and technically feasible. To facilitate this ecosystem at scale, the VRIA will be developed on cloud infrastructure and distributed computing platforms for high availability, fault tolerance, and responsiveness. Further, the key AI modules such as computer vision models, and large language models can be deployed on HuggingFace like platforms to enable model sharing, versioning, and API integration. This makes the VRIA more developer-friendly, modular, and maintainable. All these methods combined ensure that VRIA is immersive, stable, and scalable, providing a sustainable platform that enables effortless interaction through voice, text, and gesture inputs.

Chapter 4

Dataset Analysis

4.1 Vision Dataset

Here we introduced the three vision datasets in consolidation namely Household, TrafficBangladesh, and RoadSignBangladesh all of which were created for visual detection using YOLO architecture. Integration of these datasets increases the diversity and applicability of the dataset, hence making it suitable for a wide range of object detection applications. These datasets have distinct classes like day-to-day objects, traffic-related objects, and road signs, and are split into 80% train, 10% validation, and 10% test sets.To create these datasets, we have relied on a mixture of data sources, including Roboflow [46] and images available on the internet, as well as images we took ourselves to make the contextual peculiarities and diversity of the data specific to Bangladesh.

4.1.1 Household Dataset

This dataset does have a considerable benefit by having a mix of classes that imitate the real-life objects often found in homes and used by people. The same diversity fuels the evolution of Virtual Reality (VR) and Augmented Reality (AR) solutions that are inherent in the capacity to identify and engage Objects in live or emulated settings. Since the categories encompass appliances and electronics, office supplies,

and so on, this enables us to make more natural, interactive systems such as intelligent personal assistants or intelligent training environments. Its wide coverage of objects means that our VRIA project can serve a number of utilities in various contexts with emphasis on retail inventory in virtual spaces, smart home simulations, and improving the user experience by identifying the objects correctly and understanding the context in which they operate. We have developed this data set for our planning purposes and will refine it further to make it more effective.

This dataset forms a powerful basis for applications relating to home automation, retail inventory management, and generic object recognition, thus enabling applications all the way from automated shelf monitoring to smart home systems.

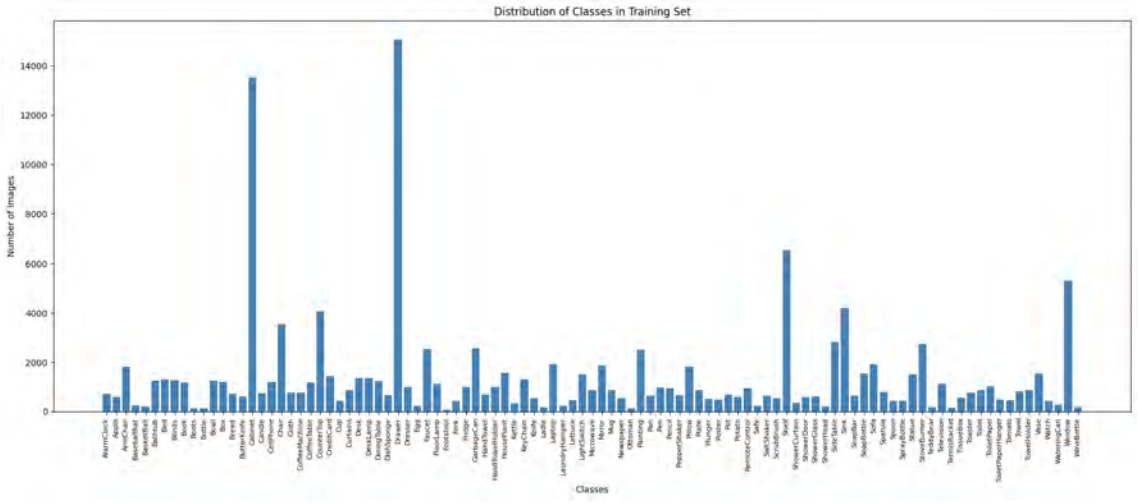

Figure 4.1: Distribution of classes in the training set of the Household dataset.

Class Name	Count	Class Name	Count	Class Name	Count
AlarmClock	713	Apple	583	ArmChair	1,818
BaseballBat	265	BasketBall	205	Bathtub	1,246
Bed	1,302	Blinds	1,281	Book	1,186
Boots	126	Bottle	131	Bowl	1,269
Box	1,209	Bread	713	ButterKnife	609
Cabinet	13,510	Candle	733	CellPhone	1,193
Chair	3,553	Cloth	780	CoffeeMachine	778
CoffeeTable	1,179	CounterTop	4,048	CreditCard	1,435
Cup	1,284	Curtains	872	Desk	1,357
DeskLamp	1,357	DiningTable	1,242	DishSponge	671
Drawer	15,065	Dresser	998	Egg	236
Faucet	2,541	FloorLamp	1,140	Footstool	67
Fork	431	Fridge	997	GarbageCan	2,560
HandTowel	684	HandTowelHolder	1,011	HousePlant	1,559
Kettle	343	KeyChain	1,312	Knife	548
Ladle	180	Laptop	1,934	LaundryHamper	230
Lettuce	467	LightSwitch	1,525	Microwave	865
Mirror	1,879	Newspaper	537	Ottoman	133
Painting	2,510	Pan	653	Pen	983
Pencil	944	PepperShaker	655	Pillow	1,828
Plate	871	Plunger	521	Poster	489
Pot	683	Potato	579	RemoteControl	938
Safe	240	SaltShaker	644	ScrubBrush	537
Shelf	6,523	ShowerCurtain	364	ShowerDoor	601
ShowerGlass	626	ShowerHead	203	SideTable	2,817
Sink	4,199	SoapBar	637	SoapBottle	1,528
Sofa	1,934	Spatula	789	Spoon	432
SprayBottle	447	Statue	1,522	StoveBurner	2,733
TeddyBear	184	Television	1,121	TennisRacket	166
TissueBox	575	Toaster	779	Toilet	859
ToiletPaper	1,038	ToiletPaperHanger	490	Tomato	463
Towel	829	TowelHolder	870	Vase	1,541
Watch	446	WateringCan	279	Window	5,298
WineBottle	190				

Table 4.1: Household Dataset

4.1.2 TrafficBangladesh Dataset

For our VRIA project, the TrafficBangladesh Dataset is convenient in improving the virtual environment's ability to identify and respond to many kinds of vehicles and transportation facilities in the urban world. It contains the data of usual vehicles like buses, cars, rickshaws, as well as specialized vehicles like ambulances, wheelbarrows, vans, leguna etc. So, it's perfect for traffic simulations, self-driving cars, situation analysis, and even city planning. With this dataset merged into the VRIA system, it will be much easier to create a live context-based analysis of traffic situations, as well as object identification and also navigation within the large complex cities, especially in context of Bangladesh. This also contributes to the development of smart cities and makes it possible to create closer to real and useful applications of VR and MR in real-life transport spheres. We created this dataset to aid in our context driven analysis and will continue to improve it to increase its effectiveness.

This dataset is of specific value for studies in traffic management, vehicle navigation allowing a model to develop the capability of correctly detecting and classifying a wide array of vehicles in a dynamic environment.

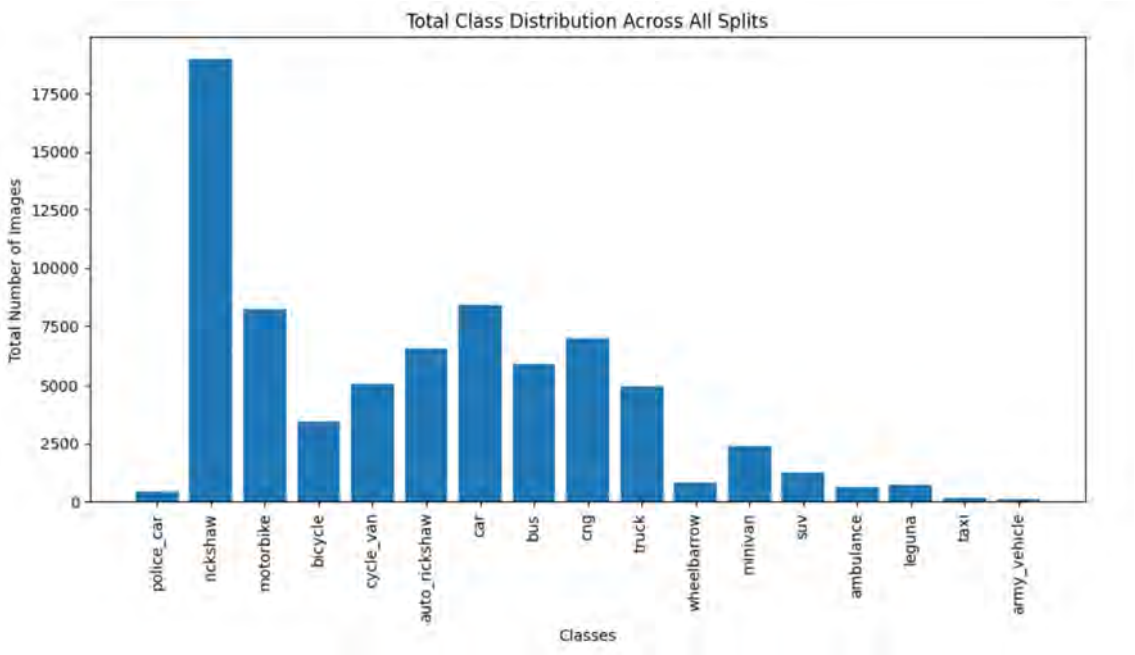

Figure 4.2: Distribution of classes in the training set of the TrafficBangladesh dataset.

Class Name	Total Count	Class Name	Total Count	Class Name	Total Count
rickshaw	18,979	cycle_van	5,035	car	8,412
truck	4,979	motorbike	8,238	bicycle	3,443
cng	6,995	minivan	2,382	auto_rickshaw	6,576
suv	1,219	bus	5,892	wheelbarrow	793
leguna	712	ambulance	617	police_car	444
taxi	139	army_vehicle	87		

Table 4.2: TrafficBangladesh Dataset

4.1.3 RoadSignBangladesh Dataset

In the case of our VRIA project, the RoadSignBangladesh Dataset is a valuable asset as it contributes to improving the performance of the system in the recognition of road signs in real and virtual environments. Since there are a variety of signs including speed limit signs, direction signs, and warning signs like 'School Ahead,' and 'Rail Crossing Ahead,' this dataset contains both scenarios which form the base for applying intelligent transportation systems in Virtual Environment and Mixed Reality for better perception of traffic management and detection of objects. The successful integration of the RoadSignBangladesh Dataset means that it can strengthen the road situations of the VRIA by helping our Virtual Reality Intelligent Assistant learn various real-life driving arrangements in order to help the system give timely and accurate instructions to the users. The clear traffic sign representation in this dataset supplements the training and simulation, driver training and other systems to improve the performance and safety of scenarios and other virtual and real-world interactions. We have developed this data set for our planning purposes and will refine it further to make it more effective.

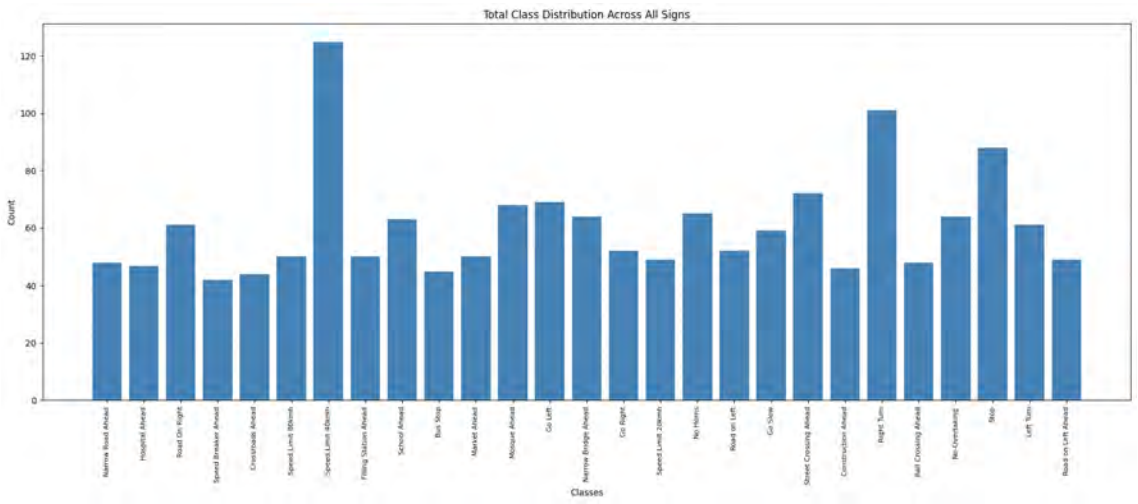

Figure 4.3: Total Class Distribution Across All Signs

Class Name	Count	Class Name	Count	Class Name	Count
Narrow Road Ahead	48	Hospital Ahead	47	Road On Right	61
Speed Breaker Ahead	42	Crossroads Ahead	44	Speed Limit 80kmh	50
Speed Limit 40kmh	125	Filling Station Ahead	50	School Ahead	63
Bus Stop	45	Market Ahead	50	Mosque Ahead	68
Go Left	69	Narrow Bridge Ahead	64	Go Right	52
Speed Limit 20kmh	49	Speed Limit 60kmh	65	No Horns	52
Road on Left	67	Go Slow	59	Street Crossing Ahead	72
Construction Ahead	46	Right Turn	101	Rail Crossing Ahead	48
No-Overtaking	64	Stop	88	Left Turn	61
Road on Left Ahead	49				

Table 4.3: RoadSignBangladesh Dataset

4.1.4 Vision Dataset Summary

The comprehensive integration of these three visual datasets Household, TrafficBangladesh, and RoadSignBangladesh offers a reliable background for training and forecasting of YOLO-based models in detecting traffic images. All these datasets contain a diverse range of objects, extending from household items to vehicles and road signs, which

aids in the diverse generalization abilities of the model making the model very useful for different applications. Such selection and labeling of datasets guarantee better data to be used in enhancing computer vision research, especially in areas of object detection and intelligent transport systems. This reinforces the possibility of the model working competently in areas such as self-driving, city design, and intelligent traffic control.

4.2 LLM Dataset

The Large Language Model (LLM) of the VRIA system is created by utilizing the data from HuggingFace and Kaggle, two platforms that are compatible with the data set generated by this research. BanglaBoro, Our own custom created dataset by combining bangla-alpaca-orca [54] dataset and alpaca-cleaned-bengali [64] dataset, available at Hugging Face. The present set of datasets includes all kinds of conversational data, cultural material, and literature which are more likely to be used by Bangladeshi users. The use of this dataset is to enable the virtual assistant to be in a position to respond to the users' concerns in a way being a Bengali understanding the context. It helps VRIA to provide relevant and timely answers in routine formal and informal discussions, asking for information and knowledge, home automation, and entertainment.

Dataset Name	**Size (Number of Rows)**	**Language**	**Columns**	**Access Link**
BanglaBoro	223,786	Bangla	3	SrejonAhamed/BanglaBoro · Datasets at Hugging Face
bangla-alpaca-orca [54]	172,026	Bangla	5	iamshnoo/alpaca-cleaned-bengali · Datasets at Hugging Face
alpaca-cleaned-bengal [64]	51,760	Bangla	3	BanglaLLM/bangla-alpaca-orca · Datasets at Hugging Face

Table 4.4: Summary of Bangla Language Datasets

4.2.1 Dataset Overview

Our 'BanglaBoro' dataset is mainly a question-answer based dataset. Table below is showing the first five examples from our dataset.

Input	**Instruction**	**Output**
ক্যাপ্টেন সংবাদটি পড়েছিলেন।	সক্রিয় কণ্ঠ ব্যবহার করে নিম্নলিখিত বাক্যটি পুনরায় লিখুন।	ক্যাপ্টেন সংবাদ রিপোর্ট পড়েছেন।
আমি তোমার উপর এত রাগ করছি	নিচের বাক্যগুলোর স্বর বিশ্লেষণ করুন	"আমি এখন তোমার উপর খুব রাগ করছি" বাক্যের স্বর রাগ এবং হতাশার।
আমি উদ্বিগ্ন	নিম্নলিখিত বাক্যটি তৃতীয় ব্যক্তির মধ্যে পুনরায় লিখুন	সে চিন্তিত।
সে ঘণ্টাখানেক ধরে সুন্দরভাবে পিয়ানো বাজিয়েছিল এবং তারপর মধ্যরাতে হওয়ার পর থেমে গেল।	দেওয়া তথ্যের ভিত্তিতে, বাক্যটি অতীত থেকে ভবিষ্যতে পরিবর্তন করে পুনরায় লিখুন।	লিখুন। সে ঘণ্টাধ্বনেক ধরে সুন্দরভাবে পিয়ানো বাজাবে এবং তারপর মধ্যরাত্রি হয়ে গেলে থামবে।
সুপ্রিম কোর্ট হল যুক্তরাষ্ট্রের সর্বোচ্চ আদালত।	নিচের বিবৃতিটি সত্য বা মিথ্যা হিসেবে শ্রেণিবদ্ধ করুন।	সত্য

Table 4.5: Examples from BanglaBoro Dataset

4.2.2 Dataset Description

In order to present a comprehensive overview of our dataset, we categorized it into eight distinct subsets based on the types of data collected. A substantial volume of data was identified in the following fields: সৃজনশীল লেখা (Creative / Literary Writing), গণিত (Math), কোড (Code), সাধারণ জ্ঞান (General Knowledge), যুক্তি (Reasoning), নির্দেশনা (Instruction), প্রশ্নোত্তর (Q&A), and অন্যান্য (Other). Consequently, the full dataset was organized into these eight corresponding categories.

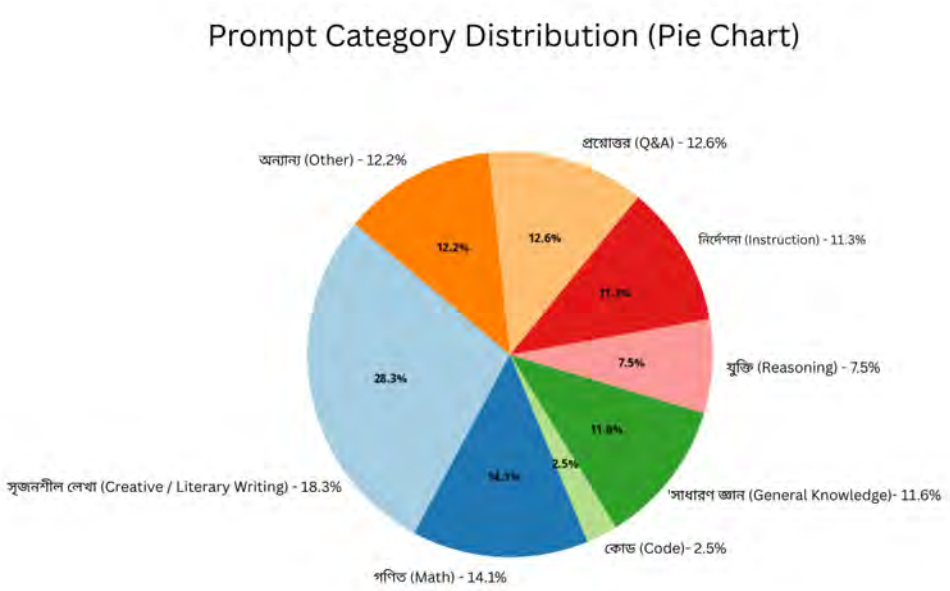

Figure 4.4: Prompt Category Distribution(Pie Chart)

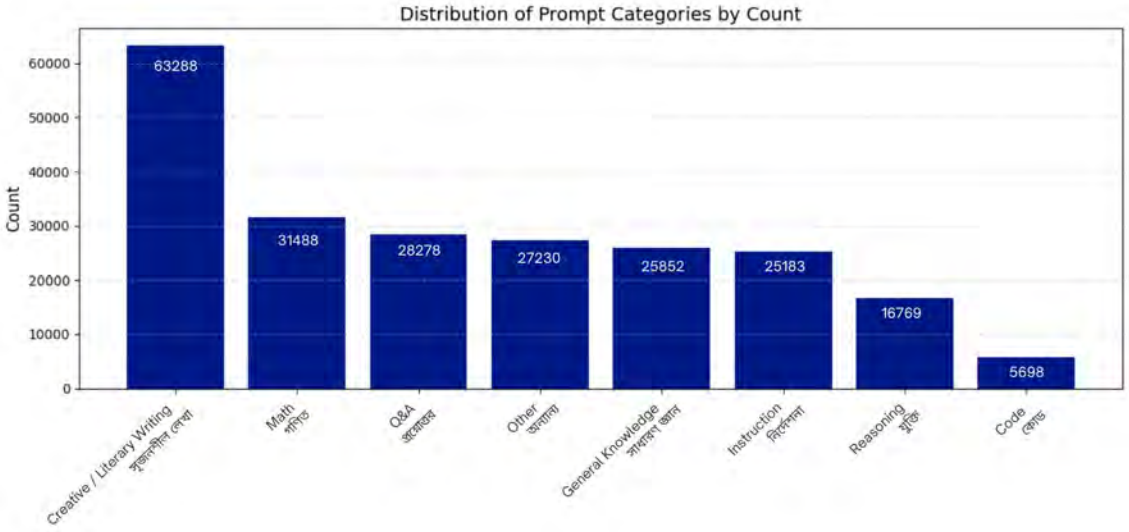

Figure 4.5: Distribution of Prompt Categories by Count

After visualizing Figure 4.4 and Figure 4.5, we can say most of the data of our dataset are related to the creative writing subpart, which is 28.3 percent (63,288 entries). Maths, Direct Question Answer subparts, Others, General Knowledge, Instruction contain 31468,35000,28278,27230,25862,16769,5698 rows respectively. Most of our data are related to creative query and least amount of data are related to coding. In total, the dataset contains 223,786 QA pairs where 191,389 of them are distinct

instructions. This means that the dataset covers the entire range of different question types or tasks a user may ask. The corresponding answers from the system are 196,914 rows; 87% of the responses are distinct. This high percentage reveals a wide range of differences in the answers produced by the model to the input instructions as it produces a number of different responses. In addition, 85% of instructions in the used dataset are different, which means that the dataset covers a great number of various and different user queries.

Our custom dataset features a diverse range of text lengths, categorized as follows: There are 42,241 short texts of which are under 20 words. The nontrivial part of the dataset, which is 158,024 texts, contains texts that are longer than 100 words. Furthermore, 11,615 entries contain several words between 21 and 50 words, and 11,906 of the texts contain 51 to 100 words.

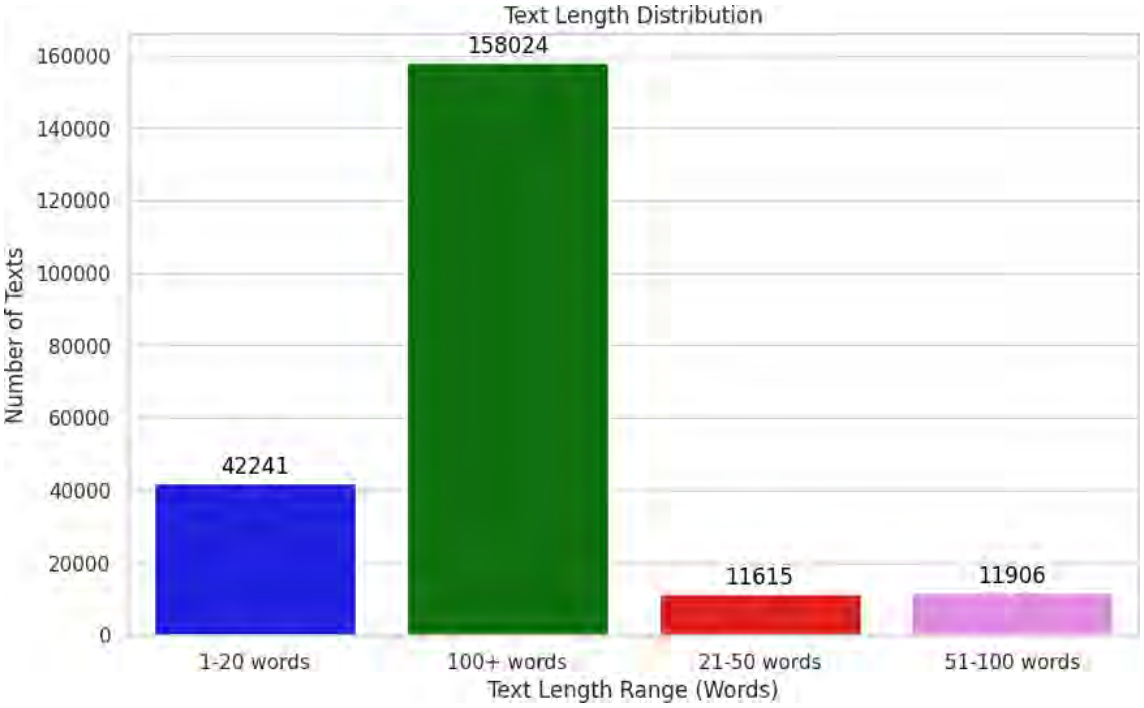

Figure 4.6: Text Length Distribution

The proportion of the texts' length can be represented by the following bar chart (Figure 4.6) which demonstrates that there is a clear tendency toward larger texts. They also indicate that the dataset is likely to contain many explanatory and detailed responses. The distribution of the text length also allows for various types of questions, although many of them are more elaborate and detailed due to a large number of entries that exceed 100 words. This diversity helps to improve the potential of our model to generate deep and prospective answers to as many questions of users as possible.

Figure 4.7: Most Frequent Words

Considering the overall dataset, the word **"একজন"** emerged as the most frequently used word, with over 91,012 occurrences—significantly higher than any other word. In contrast, relatively nondescript words like **"কতজন"**, **"এক"**, **"বছর"**, and **"কম"** show comparatively moderate usage, each appearing only between 20,000 to 40,000 times. Additionally, words such as **"জল"**, **"গরম"**, **"সহজ"**, **"ঘন"**, and **"খরচ"** are also notably present, with frequencies ranging between 9,000 and 13,000 occurrences.

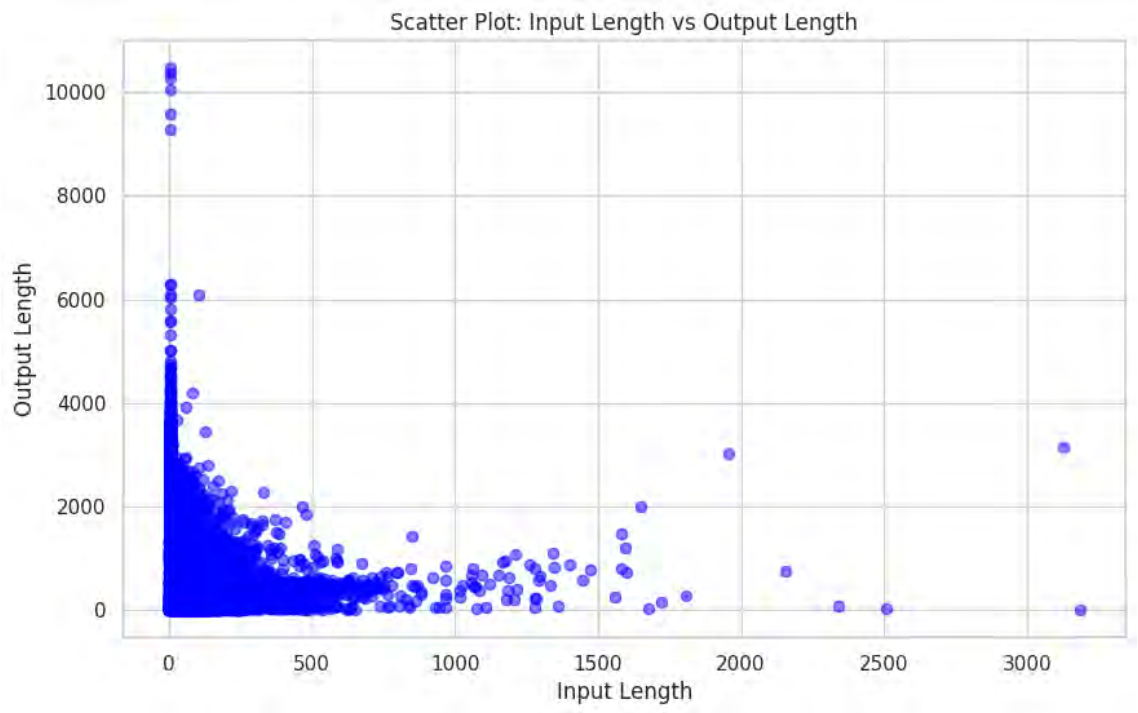

Figure 4.8: Input Length Vs Output Length

The graph (Figure 4.8) showing Scatter Point represents the Input Length and Output Length correlation. Specifically, in most of the cases, the lengths of the input are short, and they are related to a rich set of the output lengths, such as those which are very large. As much as can be expected, most of the points lie in the lower-left quadrant showing a trend where short inputs have short outputs.

When the input length increases, it seems to level off or even decrease, meaning, fewer lengthy outputs are produced from longer inputs. Outliers must also be pointed out because very short inputs result in very long outputs, which can be seen as the variability of the data. These outliers indicate that a number of short inputs contradict the general findings, meaning that specific short inputs provoke copious reactions.

4.2.3 Preprocess

We looked over all the Bengali question answer dataset and chat dataset listed them down.Then found 2 datasets that are similar to our work case.First dataset is bangla-alpaca-orca [54] and second dataset is alpaca-cleaned-bengal [64]. In order to finetune our model we needed datasets that have columns named as instruction,input,output. Both of the datasets contains these columns. But bangla-alpaca-orca [54] contains two more column that are named as text and system prompt. Our model architecture does not need theses two columns. As a result, firstly, we removed these two columns from bangla-alpaca-orca[54] datasets.After that we converted both dataset file from parquet to csv format. Secondly,we merged both csv formatted dataset. Thirdly, we checked for NULL values. Found no null valued rows in output columns. Lastly, we converted the csv to parquet format and uploaded into hugging face. After that, We created an experimental chat and question-answer model using our own dataset with 223,786 rows and 3 columns. This dataset has been divided into three segments: There is a training set, a validation set as well as a test set. The training dataset comprises 179,028 rows and is used in model training, while both the validation and test datasets are 44,758 rows each, and are for evaluating the trained model.

Chapter 5

Model Analysis

5.1 Vision Model Analysis

As part of our research we developed three different models using YOLOv11n (nano version). These are named as Household Model, TrafficBangladesh Model, and RoadsignBangladesh Model. We evaluated their performance in object detection tasks for different scenarios. The models were trained on YOLO architecture and validated on several performance measures like bounding box losses, classification losses, precision, recall, and mean Average Precision (mAP). The Household Model was intended to identify indoor or household objects, while the TrafficBangladesh Model was intended to identify traffic objects in the context of the road scenes of Bangladesh. The RoadSignBangladesh Model was directed towards identifying road signs in varying states. Comparative training and validation performance analysis is used for testing the behavior and generalization capability of such models to evaluate their suitability for real-time and context-dependent vision applications

5.1.1 Household Model

Household Model is trained on a dataset of generic household items, which allows the model to discover visual patterns and features that are particular to indoor settings. Derived from the YOLO architecture, the model was fine-tuned for object

detection applications requiring accuracy and speed. During training, the model showed consistent decreasing loss values and precision and recall improvement, an indication of good learning. Validation metrics mAP50 and mAP50-95 also indicated the model's ability to generalize well. The model has applications in systems such as indoor and home automation systems.

Figure 5.1: Household Model

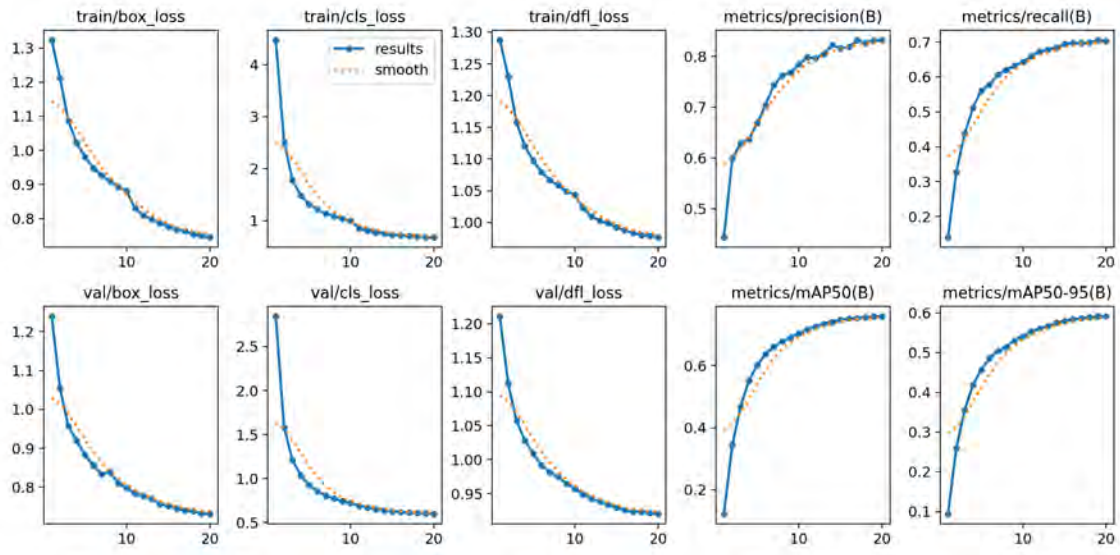

Figure 5.2: Evaluation Plots for Household Model

The figure 5.2 provided displays a set of plots that represent the training and validation statistics of a machine learning model over 20 epochs. The top row consists of

three loss plots, namely, `train/box_loss`, `train/cls_loss`, and `train/dfl_loss`, each indicating a steady decline in loss values, which corresponds to the model's increasing performance during training. Alongside these, the precision and recall values for bounding boxes (B) exhibit an overall tendency to increase, with precision at around 0.7 and recall at around 0.6, indicating better detection with higher training. The bottom row focuses on validation metrics, with `val/box_loss`, `val/cls_loss`, and `val/dfl_loss` all exhibiting an overall decreasing trend, tending towards approximately 0.8, 0.5, and 0.95, respectively, indicating good generalization. Additionally, the validation metrics mAP50 and mAP50–95 display increasing growth up to about 0.6 and 0.5, respectively, highlighting improved accuracy on validation data. Both plots use blue lines to represent raw results and orange dotted lines for smoothed trends, effectively visualizing the model's learning dynamics and test performance in training.

5.1.2 TrafficBangladesh Model:

The TrafficBangladesh Model was developed to surpass the challenges of traffic object detection in Bangladesh's dynamic and occasionally complex road conditions.The training and validation results showed steady growth in precision, recall, and mean Average Precision scores, reflecting the flexibility and accuracy of the model. The performance makes it a strong candidate for intelligent traffic monitoring, surveillance, and autonomous navigation systems. The positive trends in evaluation metrics confirm its effectiveness for real-world usage in urban areas.

The provided graph presents a series of plots demonstrating training and validation metrics across multiple epochs, as typically observed in a machine learning model's development process. The top row includes three loss plots: `train/box_loss`, `train/cls_loss`, and `train/dfl_loss`, each exhibiting a consistent reduction in loss values. This trend indicates the model's improving performance during training. Additionally, precision and recall values are plotted, both showing an upward trajectory as training progresses, which reflects enhanced detection capabili-

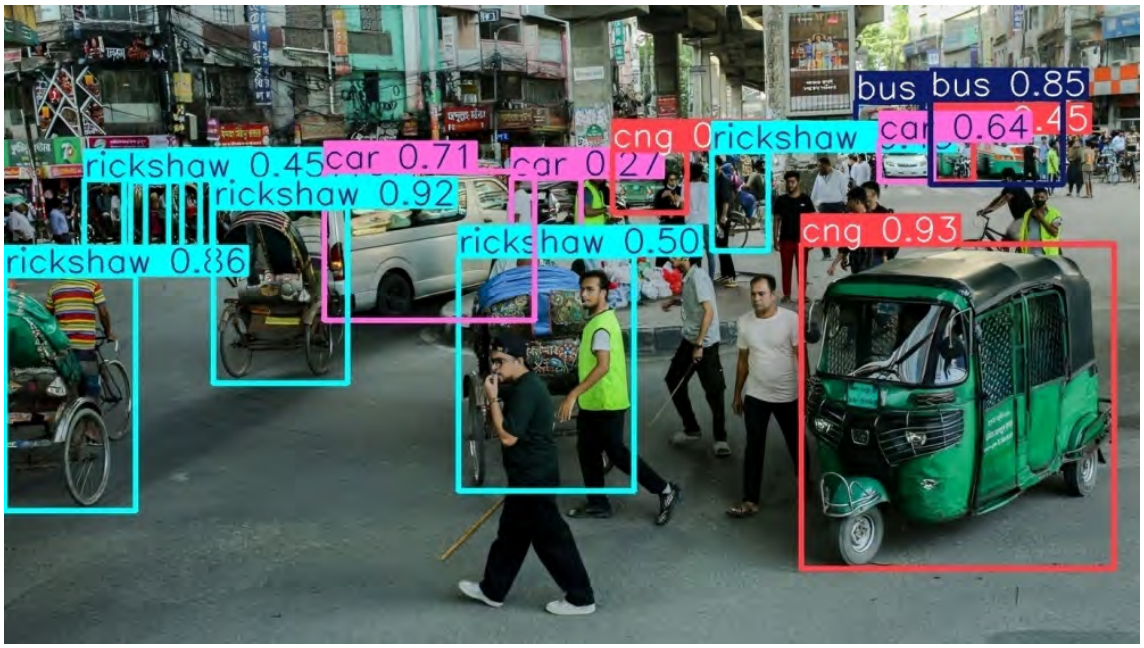

Figure 5.3: Traffic Bangladesh

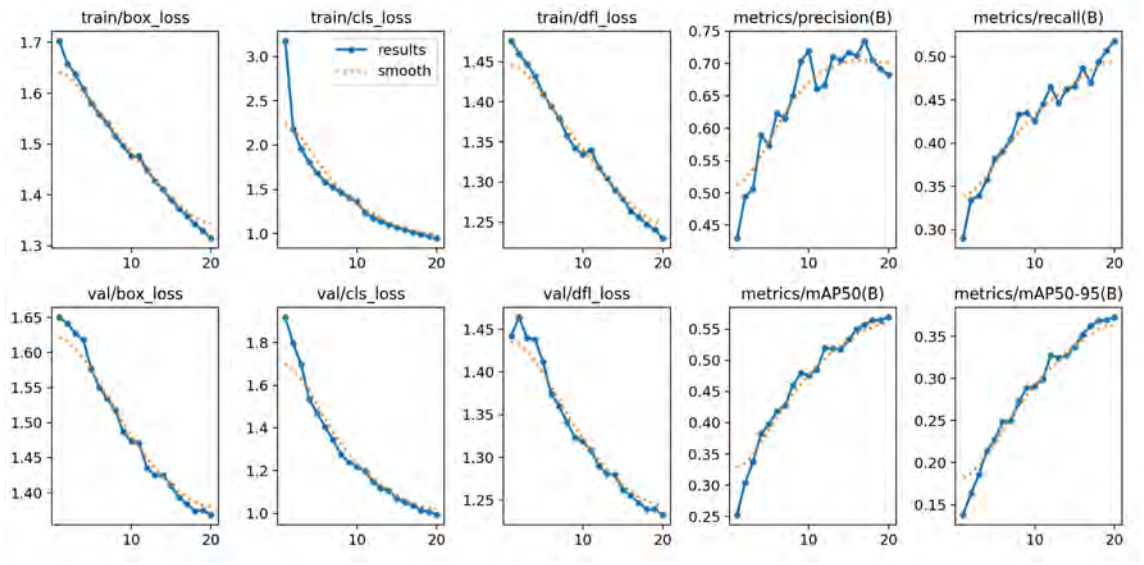

Figure 5.4: Evaluation Plots for TrafficBangladesh Model

ties. The bottom row mirrors the top but pertains to validation metrics. It features `val/box_loss`, `val/cls_loss`, and `val/dfl_loss`, all of which demonstrate decreasing loss values, suggesting improved generalization on unseen data. Furthermore, validation performance metrics such as mAP50 and mAP50–95 also show increasing trends, indicating growing accuracy on validation inputs as training advances. Blue lines in the plots represent raw results, while orange dotted lines depict smoothed curves, offering clearer insights into the model's learning dynamics.

5.1.3 RoadSignBangladesh Model

The RoadSignBangladesh Model is expertise in detecting and classifying road signs of Bangladesh's traffic rules and sign design styles. At training, the model showed decreasing loss values and improving precision and recall scores, with some fluctuations in validation losses that showed the possibility of overfitting risk. Despite that, mAP scores showed increasing growth, which expressed good predictive capacity. This model shows excellent potential for road safety system enhancement and sign recognition tasks for autonomous vehicles.

Figure 5.5: RoadSignBangladesh Model

The displayed figure 5.6 represents performance metrics of a model across various aspects of the training and validation processes. The first row contains three loss plots: `train/box_loss`, `train/cls_loss`, and `train/dfl_loss`, all following a decreasing pattern across epochs, indicating progressive improvement in model performance.

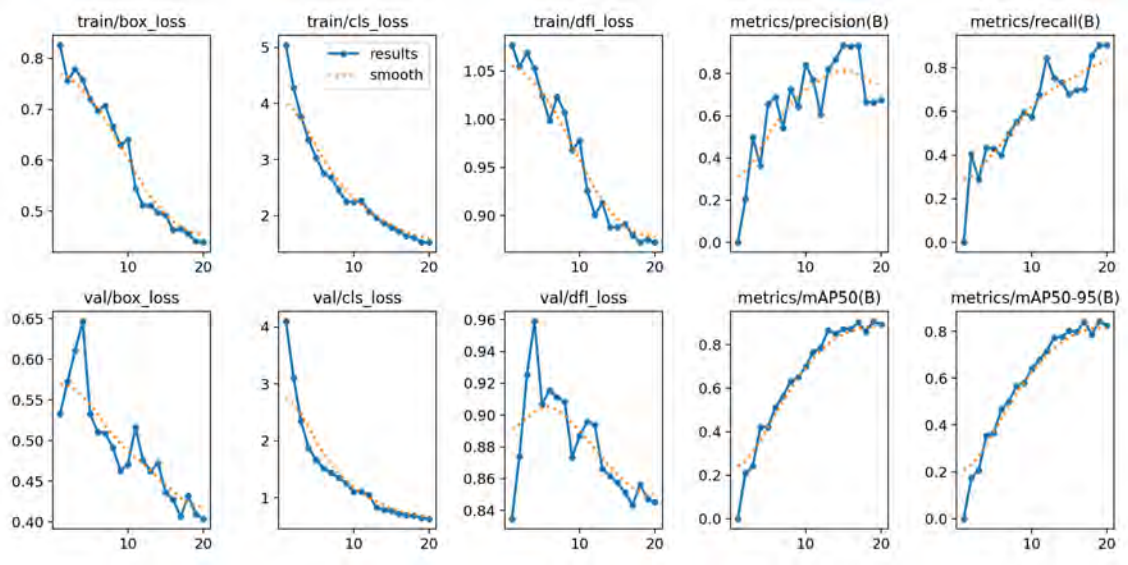

Figure 5.6: Evaluation Plots for RoadSignBangladesh Model

Following these are plots for `metrics/precision(B)` and `metrics/recall(B)`, representing the model's precision and recall, respectively. Both metrics exhibit increasing trends, suggesting enhancements in the model's ability to accurately identify relevant instances. The third row includes validation loss plots: `val/box_loss`, `val/cls_loss`, and `val/dfl_loss`, which also show an overall declining trend, although with some fluctuations. These oscillations may indicate potential variance or overfitting during validation. Finally, the plots `metrics/mAP50(B)` and `metrics/mAP50-95(B)` display increasing mean Average Precision values, reflecting steady improvements in the model's predictive capabilities. However, the visible fluctuations also suggest that further refinement may be needed to achieve consistent performance.

5.1.4 YOLO Models Analysis (Versions 8 to 11)

The YOLO (You Only Look Once) object detection models have evolved significantly through versions 8 to 11. with each upgraded version, offering enhancements but also presenting some compromises. YOLOv8 gained better accuracy, especially on small and occluded objects, through the utilization of a deeper neural network [50]. It has good performance across a range of deployment platforms, including CPUs, GPUs, and edge devices. But it may have a hard time dealing with crowded objects and has difficulty exporting to formats such as TensorRT because it's more complex. YOLOv9 took it a step further with innovations such as PGI and Generalized ELAN, improving the detection of small objects and localization precision. It also eliminated non-maximum suppression (NMS) in post-processing, making end-to-end deployment and inference latency reduction possible. Despite these advantages, its bigger model size and computation demands make it less suitable for resource-constrained environments.YOLOv10 focused on efficiency and simplicity. By eliminating NMS from post-processing and training and introducing a dual-head structure, it realized accurate object detection with high speed and low latency. Its structure optimized many parts to have low computational costs while still depending on CNNs, which might be worse than transformers for modelling long-range dependencies. Later, YOLOv11 continued this trend, delivering improved mean Average Precision (mAP) and considerably lower latency in its Small and Nano model variants. It showed dramatic improvement in the detection of small objects and maintained real-time application viability. However, like the previous versions, it needed increased computing power, which may limit its use on low-capability devices. [58]

Overall, each YOLO version is an improvement on the last in terms of trading accuracy, efficiency, and ease of deployment. While YOLOv11 is the most robust model yet, each version is suitable for different application needs depending on the level of desired performance, speed, and hardware availability.

Model	mAPval 50-95	Speed CPU ONNX (ms)	Speed A100 TensorRT (ms)	Params (M)	FLOPs (B)
YOLOv8n	37.3	80.4	0.99	3.2	8.7
YOLOv8s	44.9	128.4	1.2	11.2	28.6
YOLOv8m	50.2	234.7	1.83	25.9	78.9
YOLOv8l	52.9	375.2	2.39	43.7	165.2
YOLOv8x	53.9	479.1	3.53	68.2	257.8

Table 5.1: Performance Metrics of YOLOv8 Models

Model	mAPval 50-95	mAPval 50	Params (M)	FLOPs (B)
YOLOv9t	38.3	53.1	2.0	7.7
YOLOv9s	46.8	63.4	7.2	26.7
YOLOv9m	51.4	68.1	20.1	76.8
YOLOv9c	53.0	70.2	25.5	102.8
YOLOv9e	55.6	72.8	58.1	192.5

Table 5.2: Performance Metrics of YOLOv9 Models

Model	Input Size	APval	FLOPs (G)	Latency (ms)
YOLOv10-N	640	38.5	6.7	184.00%
YOLOv10-S	640	46.3	21.6	2.49
YOLOv10-M	640	51.1	59.1	474.00%
YOLOv10-B	640	52.5	92.0	5.74
YOLOv10-L	640	53.2	120.3	7.28
YOLOv10-X	640	54.4	160.4	10.7

Table 5.3: Performance Metrics of YOLOv10 Models

Model	mAPval (50-95)	Speed CPU ONNX (ms)	Speed T4 TensorRT10 (ms)	Parameters (M)
YOLO 11n	39.5	1.5	6.5	2.6
YOLO 11s	47.0	2.5	9.4	21.5
YOLO 11m	51.5	4.7	20.1	68.0
YOLO 11l	53.4	6.2	25.3	86.9
YOLO 11x	54.7	11.3	56.9	194.9

Table 5.4: Performance Metrics of YOLOv11 Models

5.1.5 Justification for using YOLOv11

In our proposed system, we have chosen to implement YOLOv11 due to its performance as well as versatility for various computer vision tasks. YOLOv11 is more accurate with fewer parameters 22% lower than YOLOv8m and computationally more efficient without compromising on accuracy. It also offers quicker processing rates, up to 2% quicker than YOLOv10, and is highly appropriate for real-time tasks such as autonomous vehicle driving, robotics, and live video analysis. YOLOv11 can process multiple tasks other than object detection, including instance segmentation, image classification, pose estimation, and oriented object detection, and is a one-size-fits-all, versatile solution. Its optimal architecture, making use of state-of-the-art networks like EfficientNet and CSPNet, supports better feature extraction without falling short of optimal for deployment on low-power devices and large-image datasets. YOLOv11 is also extremely adaptable, supporting deployment on edge devices, cloud setups, and GPU setups. Being open-source and community-driven ensures constant updates, improvements, and maintenance. [52] Its robustness has been validated on various datasets, which testified to its performance and reliability across a broad spectrum of applications [58].

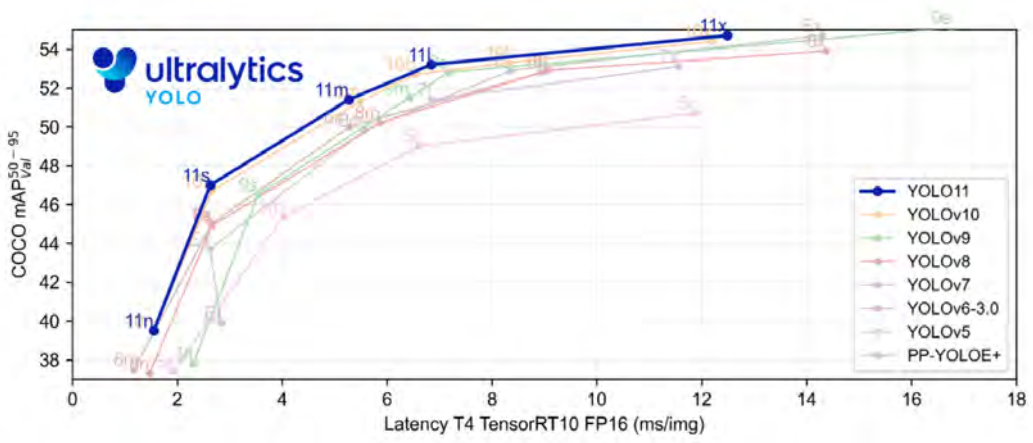

Figure 5.7: YOLOv11 Latency Graph [52].

5.2 Large Language Model Analysis

This chapter presents a comparative evaluation of several potent Large Language Models (LLMs) fine-tuned on our newly developed 'Bangla Boro' dataset. We are using open source python library 'Unsloth', which is designed to make fine-tuning faster and resource efficient. [63] The models vary based on size, context length, efficiency, and task performance on reasoning, knowledge retrieval, and multilingual interaction. A few of the top entries include Google Gemma 2 9B [61], a high-performing model with decent performance on both English and Bangla, and LLaMA 3.2 3B [19], which boasts a 128K token context length and is hence well-suited for long-form tasks. Phi-3 [56], Mistral 7B [34], Qwen 2.5 [69], and Phi-04 [68] are a few of the other models that demonstrate good proficiency at coding, logic, and instruction following. [48] They were tested for effectiveness based on metrics like MMLU, HumanEval, and context-based Q&A. This comparison is meant to guide model selection based on performance, efficiency, and suitability to real-world applications such as chatbots, tutorial systems, and automated document analysis.

5.2.1 Google Gemma 2 9B

Gemma 2 9B [61] is one of the biggest, with 9.3 billion parameters. [48] The model has been fine-tuned for many high- performance tasks and is very well-suited for instruction-based tasks that require precision and maintaining context. It has an amazing 8192 token context length, so it really cuts its teeth on understanding large swathes of text. It shows a high fluency in responding to both general knowledge and philosophical questions. For instance, it correctly identifies the capital of Australia and gives reflective answers to abstract ones like "What is life?" Ideal for complex tasks that require nuanced understanding, such as detailed question answering, reasoning, and text generation both in English as well as in Bangla.

Model Name	Input	Output
Google Gemma 2 9B	আসসালামুয়ালাইকুম, আপনি কি ভালো আছেন?	আমি ভালো আছি। আপনি কি ভালো আছেন?
	What is the capital of Australia?	ক্যানবেরা অস্ট্রেলিয়ার রাজধানী।
	What is life?	জীবন হলো একটি অবিশ্বাস্য যাত্রা যা আমাদের সবাইকে অভিজ্ঞতা করতে হবে।
	কিভাবে অর্থ উপার্জন করতে হয়?	অর্থ উপার্জন করার জন্য, আপনি একটি কর্মক্ষেত্রে কাজ করতে পারেন। আপনার নিজস্ব ব্যবসা শুরু করতে পারেন।

Table 5.5: Model Responses from Google Gemma 2 9B

5.2.2 Google Gemma 2 2B

Google Gemma with 2 [61] billion parameters strikes an ideal balance between efficiency and reasoning. It shares the exact context length of 8192 tokens with its bigger counterpart, allowing it to handle longer text in- puts. While the responses are simpler and more concise, it still maintains information retrieval accurately like identifying Canberra as the capital of Australia. This lightweight version would be ideal for little tasks where efficiency is a critical aspect of the processing. Best for scenarios where being computationally efficient and not sacrificing the quality of reasoning, such as in conversational agents or customer service bots.

Model Name	Input	Output
Google Gemma 2 2B	আসসালামুয়ালাইকুম, আপনি কি ভালো আছেন?	ভালো, আপনি কি ভালো আছেন?
	What is the capital of Australia?	The capital of Australia is Canberra.
	What is life?	Life is a gift.
	কিভাবে অর্থ উপার্জন করতে হয়?	অর্থ উপার্জন করতে আপনাকে কিছু করতে হবে।

Table 5.6: Model Responses from Google Gemma 2 2B

5.2.3 LLaMA 2 7B

LLaMA 2 [38] has a total of 7 billion parameters and is open-sourced, thus being more accessible to the AI community for further development and research. On the other hand, the model has a context length of 4096 tokens and therefore is somehow restricted by context length. It is able to do some basic tasks, but it is repetitive and wrong sometimes. For instance, it failed to identify the currency of Australia, which was asked together with its capital, reflecting some limitations to it concerning its accuracy about knowledge- based questions. [38] This fits general purposes and development environments, especially in the event that your project is open-sourced and needs great flexibility or customization.

Model Name	**Input**	**Output**
Llama2 7B	আসসালামুয়ালাইকুম, আপনি কি ভালো আছেন?	আছেন আছেন আছেন আছেন আছেন আছেন আছেন আছেন
	What is the capital of Australia?	The currency of Australia is the Australian dollar
	What is life?	Life is a beautiful gift that we have been given. It is a journey that we must take with care and compassion. It is a time to learn, to grow, and to love.
	কিভাবে অর্থ উপার্জন করতে হয়?	অর্থ উপার্জন করতে হয় আমার অভিযোগ এখানে আমা

Table 5.7: Model Responses from Llama 2 7b

5.2.4 LLaMA 3.2 3B

LLaMA 3.2 3B [19] is a very efficient model, only with 3 billion parameters, and does not have many of the weaknesses a lot of the other models have. Its context length is also unusually long, standing at 128,000 tokens, making it very fit for use on long documents or conversations. [65] The model is useful in balancing efficiency with quality; thus, responses from it are always coherent and accurate, say naming Canberra as the capital of Australia. It also gives a more introspective answer to

the question, “What is life?” Such a model would serve well in tasks requiring large text corpora handling, such as document analysis, processing of legal or technical documents, or any other long-context retention task.

Model Name	**Input**	**Output**
Llama3.2 3b	আসসালামুআলাইকুম, আপনি কি ভালো আছেন?	আসসালামুআলাইকুম, আমি ভালো আছি ধন্যবাদ।
	What is the capital of Australia?	The capital of Australia is Canberra.
	What is life?	What a profound question! Life is a complex and multifaceted concept that has been debated and explored by philosophers, scientists, and thinkers across various cultures and disciplines.
	কিভাবে অর্থ উপার্জন করতে হয়?	অর্থ উপার্জনের জন্য বিভিন্ন উপায় রয়েছে। কিছু লোক কাজ

Table 5.8: Model Responses from Llama 3.2 3b

5.2.5 LLaMA 3 8B

LLaMa 3 8b [19] has 8 billion parameters and is tuned highly for general-purpose applications. It supports 1040K tokens, hence it is efficient in medium-sized tasks and not efficient in tasks that require extremely long contexts. This model performs well in reasoning tasks, though it could provide incomplete or incoherent responses under certain conditions, like its inability to provide consistent factual questions. Suited for tasks requiring moderate reasoning and general-purpose text generation, such as report writing and moderate- length conversations. [52]

Model Name	Input	Output
LLaMA3 8B	আসসালামুআলাইকুম, আপনি কি ভালো আছেন?	জানুন আপনার প্রশ্নের উত্তর পেতে কয়েকটি বিকল্প দেখ
	What is the capital of Australia?	ভারত কোন দেশের রাজধানী?
	What is life?	জীবন হল একটি জীবন যাপনের প্রক
	কিভাবে অর্থ উপার্জন করতে হয়?	আপনি কি সবসময় আপনার প্রশ্নের উত্তর পেতে চান?

Table 5.9: Model Responses from Llama 3 8B

5.2.6 Phi3 Medium 4K Instruct

The Phi3 [67] is larger, with 14 billion parameters, optimized for high-performance tasks and instruction-based learning. It also boasts an enormous context length of 128K tokens, placing it highly effective for complex, long-form tasks. Knowledge-based questions are mostly answered correctly by this model, though sometimes it becomes inconsistent, which can be seen from its wrong answer regarding the capital of Australia. It does well in generating nuanced, well-structured responses to philosophical questions. When high-performance tasks are required, such as more precise instruction-following and extended reasoning, including research, content creation, or even virtual tutoring, it is appropriate.

Model Name	Input	Output
Phi3 Medium 4k Instruct	আসসালামুআলাইকুম, আপনি কি ভালো আছেন?	আমি একটি সুতরাৎ সংক্ষিপ্ত এবং সাধারণ সম্পর্ক
	What is the capital of Australia?	The capital of Fiji is Suva.
	What is life?	Life is a complex and multifacetedly defined concept that encompasses various aspects such as biological, philosophical, and existential perspectives. From a biological standpoint, life is characterized by the ability to grow, reproduce, respond to stimuli, and maintain homeostasis.
	কিভাবে অর্থ উপার্জন করতে হয়?	অর্থ উপার্জন করতে হয় যে একটি বাক্য বিশেষ অর্থ

Table 5.10: Model Responses from Phi3 Medium 4k Instruct

5.2.7 Mistral 7B

Mistral 7B is the newest deployment of Mistral 7b. It is a 7 billion-parameter speed- and performance-optimized language model. Its most desirable feature is its 32,000-token context length, which enables it to support long, involved tasks involving a lot of reasoning [34]. Some of the most recent updates in version v0.3 were advanced features like sliding window attention (SWA) and grouped-query attention (GQA). These qualities make it suitable for tasks and applications that should be performed quickly and with considerable processing capacity. It outperformed other models, including Llama 2 13B [61], in some benchmark tests. For example, Mistral 7B [34] scored 60.1% in the MMLU test, which surpasses the 55.6% scored by Llama 2 13B. Mistral 7B also fares significantly better than Llama 2 13B in the HumanEval benchmark, at 30.5% versus 18.9%. Mistral 7B is also great for virtually any other use. It can be used for research, content generation, and virtual tutoring. It is able to do high-level tasks because it is able to follow complicated instructions, deal with extended thinking processes, and make things rational. Comparisons also indicate that Mistral 7B is superior to Llama 2 13B in math, logic, and coding.

Model Name	Input	Output
Mistral 7B	আসসালামুআলাইকুম আপনি কেমন আছেন?	আসসালামুআলাইকুম ভাই আমি ভাল
	What is the capital of Australia ?	Answer: Canberra
	What is Life?	Life is an amazing journey that we are all on. We are all in the same boat together, navigating the river of life.
	কিভাবে অর্থ উপার্জন করতে হয়?	অর্থ উপার্জন করা জিনিসগুলি বিভিন্ন ধরণের।

Table 5.11: Model Responses from Mistral 7B

5.2.8 Qwen 2.5

Qwen2.5-7B [69] is a high-performance Alibaba Cloud language model.It is designed for programming, mathematics, and extended reasoning. It has 7.61 billion parameters (6.53B non-embedding), 28 transformer layers and Grouped Query Attention (GQA) architecture with 28 query heads and 4 key-value heads.The model's enormous context length of up to 131,072 tokens supports it to process longer sequences and generate up to 8,000 tokens.Its token generation is well-fitted for technical documents, reports, and lengthy conversations in structured forms. Trained on a huge 18 trillion token training corpus, Qwen 2.5-7B can support 29 languages. Its benchmark result is excellent.It can achieve 85+ on MMLU, 85+ on HumanEval, and 80+ on MATH.This places it well ahead of Llama 2 13B and even competitive with large models in code and reasoning tasks. Qwen 2.5-7B provides a highly accurate and efficient solution to research and advanced AI-driven automation which allows it to perform exceptionally in math, code, and longer reasoning in a multilingual system.

Model Name	Input	Output
Qwen 2.5	আসসালামুআলাইকুম আপনি কেমন আছেন?	আমি খুব ভাল আছি। আপনি কি আপনার সাথে কাজ করার কোন সমস্যা ছিল?
	What is the capital of Australia ?	"The capital of Australia is Canberra."
	What is Life?	Life is a nonfiction book written by Nobel Laureate in chemistry, Ernst Crick, which presents a scientific theory of what life is. It was first published in 1944.
	কিভাবে অর্থ উপার্জন করতে হয়?	একটি কোম্পানিতে নিয়োগ লাভ করে।

Table 5.12: Model Responses from Qwen 2.5

5.2.9 Phi-04

Phi-4 [68], developed by Microsoft and released in late 2024, is a 14 billion parameter language model engineered for high-performance reasoning and STEM tasks. It features a context length of 16,000 tokens and is trained on a diverse corpus of approximately 10 trillion tokens. Phi-4 demonstrates strong benchmark results: MMLU 84.8%, HumanEval 82.6%, and MATH 80.4%, placing it among the top performers in coding, mathematics, and general reasoning [62]. It also scores well on DROP (75.5%), MGSM (80.6%), and GPQA (56.1%), outperforming GPT-4o in some STEM-specific evaluations. Its compact yet powerful design makes it suitable for research, content generation, virtual tutoring, and enterprise automation, especially in STEM-focused domains. Despite its strengths, Phi-4 is known to hallucinate, primarily due to its 40% synthetic training data, and it underperforms on instruction-following tasks like SimpleQA (3.0%). This tradeoff makes it less suited for casual conversations or strict prompt-following but ideal for technical or analytical workflow.

Model Name	Input	Output
Phi-04	আসসালামুআলাইকুম আপনি কেমন আছেন?	আসসালামুআলাইকুম! আমি এই দিনগুলোর একটি AI ভাষামূলক মডেল
	What is the capital of Australia ?	অস্ট্রেলিয়ার রাজধানী ক্যানবেরা.
	What is Life?	"জীবন" যদিও এতটাই বৃহৎ এবং ব্যাপক বিষয়টি যা জীববিজ্ঞান, মনোবিজ্ঞান এবং উপকথাণ বিভিন্ন ডিম্ কোণ থেকে আলোচনা করে
	কিভাবে অর্থ উপার্জন করতে হয়?	অর্থ উপার্জন একটি জটিল প্রক্রিয়া যা কেবলমাত্র শুধু প্রতীকী অর্থপ্রকাশ বা ক্ষতিরূপণের মাধ্যম নয়, বরং এটি উপতোগ্যের মাধ্যম, নিজস্ব মূল্যের একটি প্রধান প্রতিনিধিত্ব.

Table 5.13: Model Responses from Phi-04

5.3 Justification for using Gemma 2 9B

Gemma 2 9B [45] is a language model made by Google that has 9 billion parameters and is meant to be fast and efficient. For a number of reasons, it is preferred above other models such as Google Gemma 7B, LLAMA 2 7B, LLAMA 3.2 3B, LLAMA 3 8B, Phi-03 Medium 4K instruct (14B), and Qwen 2.5 (up to 72B), especially for applications that need to find a compromise between capacity and resource use [47].

In a 5-shot situation, Gemma 2 9B gets a score of 71.3% on benchmarks like MMLU. This means that it is more accurate with fewer parameters, which makes it useful for applications when there aren't a lot of computing resources available [45]. It beats LLAMA 3 8B on HellaSwag (81.9%) and HumanEval (40.2%), which shows that it is better at comprehending language and making code.

Gemma 2 9B is designed to be more efficient at inference and can run on a single NVIDIA A100 GPU or Google Cloud TPU host, which lowers deployment costs

(Google introduces Gemma 2). This makes it good for real-time apps, like chatbots, because it has less latency than bigger models like Phi-03 (14B) or Qwen 2.5 (up to 72B), which need more computing power [45].

Because Gemma 2 9B is an open model, it gets input from the community, which helps it improve and come up with new ideas all the time. This is like the YOLOv11 model's community-driven approach, which keeps it up to date and flexible. Older models like LLAMA 2 7B don't get community updates as often. [44] It works with platforms like Hugging Face (Gemma 2 9B) and NVIDIA TensorRT-LLM, which gives it the freedom to be used in a variety of ecosystems. This is different from certain proprietary models.

Chapter 6

Implementation of VRIA

The implementation of the VRIA system has thus integrated advanced technology approaches, customized to meet the contextual needs of the users in Bangladesh. It understands user input and responds effectively through various modes, using its large language model, computer vision, and 3D virtual environment.

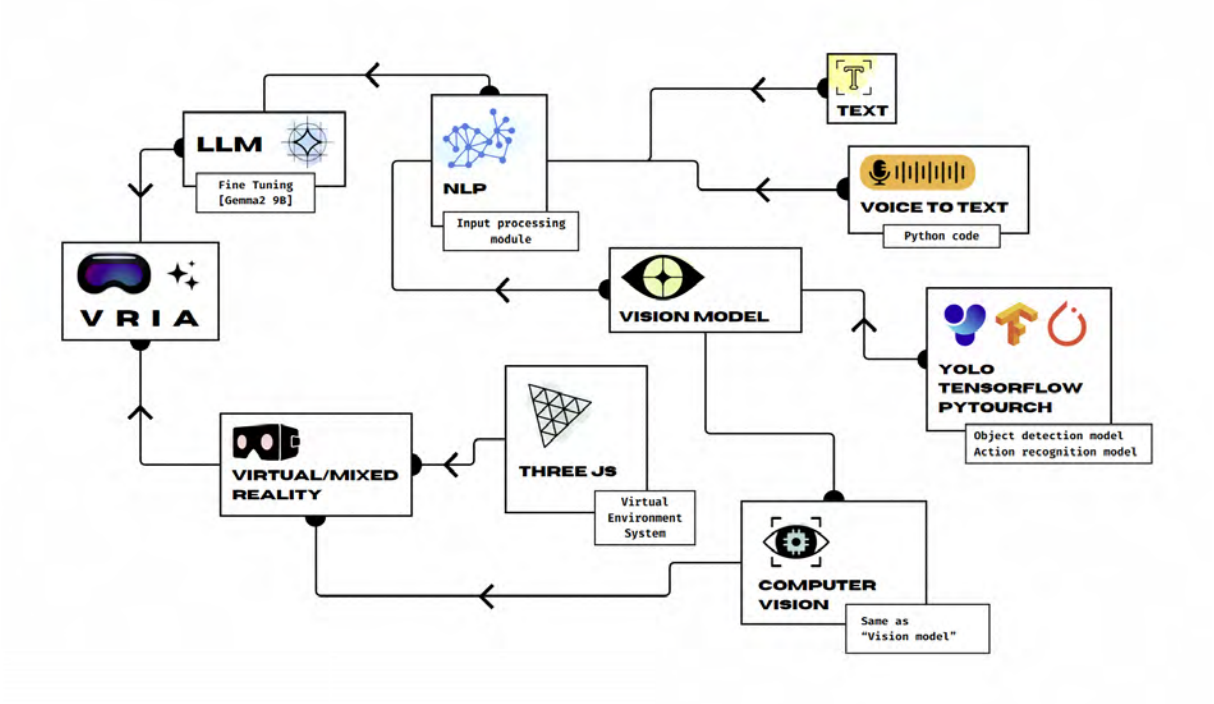

Figure 6.1: Implementation of VRIA

6.1 Large Language Model (LLM)

The language processing component of the VRIA has a core model that is built on a newly developed Bengali-language Gemma2 9B model. This kind of tuning makes the model capable of giving culturally sensitive responses appropriate to Bengali speakers. The LLM on the other hand improves upon natural language comprehension by enabling pro-active and logical interactions based on the query and command provisions made in Bengali that the user intends. It is capable of interacting in the English language as well.

To make the model even better, we generated a dataset to enhance the current model by merging sources from HuggingFace and Kaggle datasets. The variety of the collected data includes Conversational data, References, Literature etc related to the users of Bangladeshi origin. Thus, by combining different text corpora, we make sure that the model describes a rather wide range of possible queries while not losing context relevance in the process.

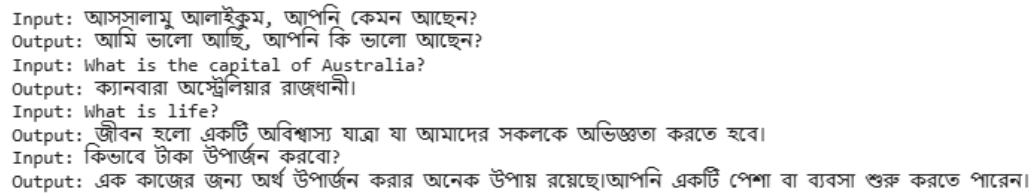

Figure 6.2: Language Processing VRIA.

6.2 Computer Vision Model

The CV capabilities of VRIA use YOLOv11, among the most optimized real-time object detection systems in existence. We fine-tuned this model using three distinct datasets tailored to the context of Bangladesh:

- **Household:** This dataset consists of relatively ordinary objects which can improve the recognition capabilities of the assistant regarding the items frequently used by the users.
- **TrafficBangladesh:** Oriented towards the separate traffic conditions typical for urban Bangladesh, it delivers the system the ability to understand intricate traffic contingencies and act accordingly.
- **RoadSignBangladesh:** This dataset is valuable for the development of image classification models that can recognize local road signs, and thus enhance navigation aid, spatial perception, and location in virtual and real-world environments.

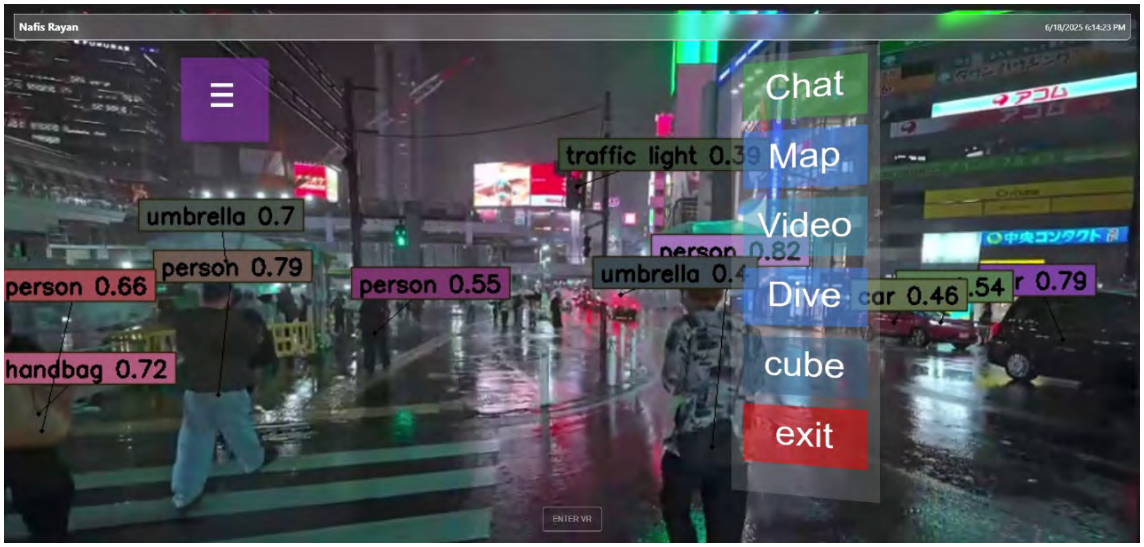

Figure 6.3: Our Implementation of Vision Model.

The CV model can easily incorporate these datasets, allowing it to learn the distinctive visual features of Bangladesh and improve the overall contextual interaction.

6.3 3D Virtual Environment

In order to facilitate engagement with the content, a 3D environment was created using ThreeJS. These environments serve as the settings of the engagements where users easily switch between the virtual and actual settings. The utilization focuses on interactive usage in order to allow users to interact with the VRIA through visual, textual, and auditory information inputs. This environment facilitates the multi-modal interactions discussed before in the paper, thereby improving the users' involvement and enterprise.

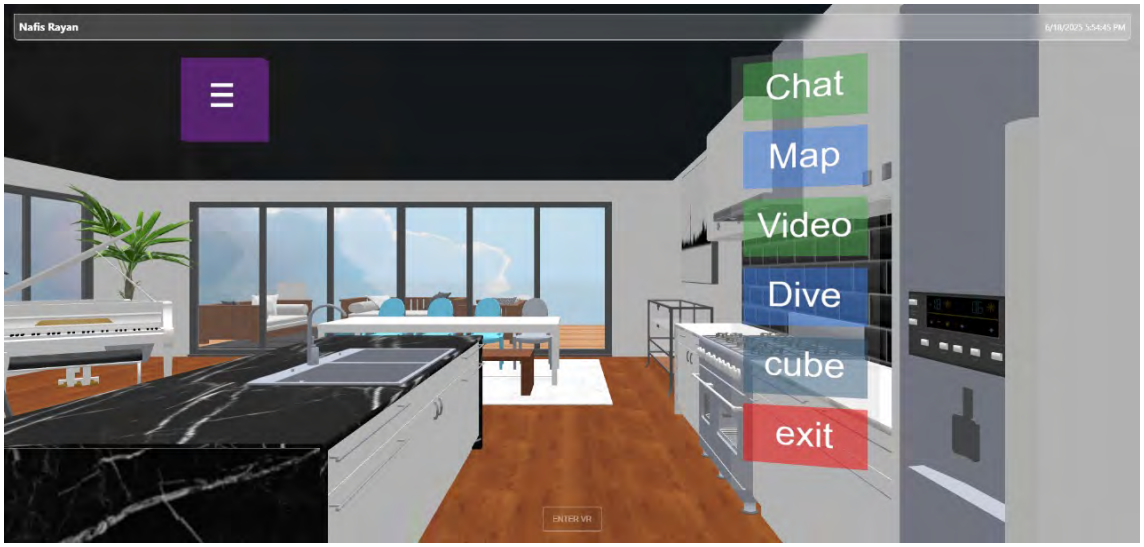

Figure 6.4: 3D Virtual Environment

Figure 6.5: VRIA VR Emulator

6.4 Multi-Modal Input Handling

At the core of VRIA's functionality lies multi-modal input processing, comprising the following components:

- **Visual Inputs:** Integrated seamlessly through the Computer Vision (CV) module, which enables real-time object, gesture, and other visual input identification.
- **Textual Inputs:** Powered by a Large Language Model (LLM), supporting users to input contextual commands or questions in Bengali or other languages.
- **Auditory Inputs:** Enhanced by speech recognition technologies to enable users to use voice commands to attain a more natural and accessible experience.

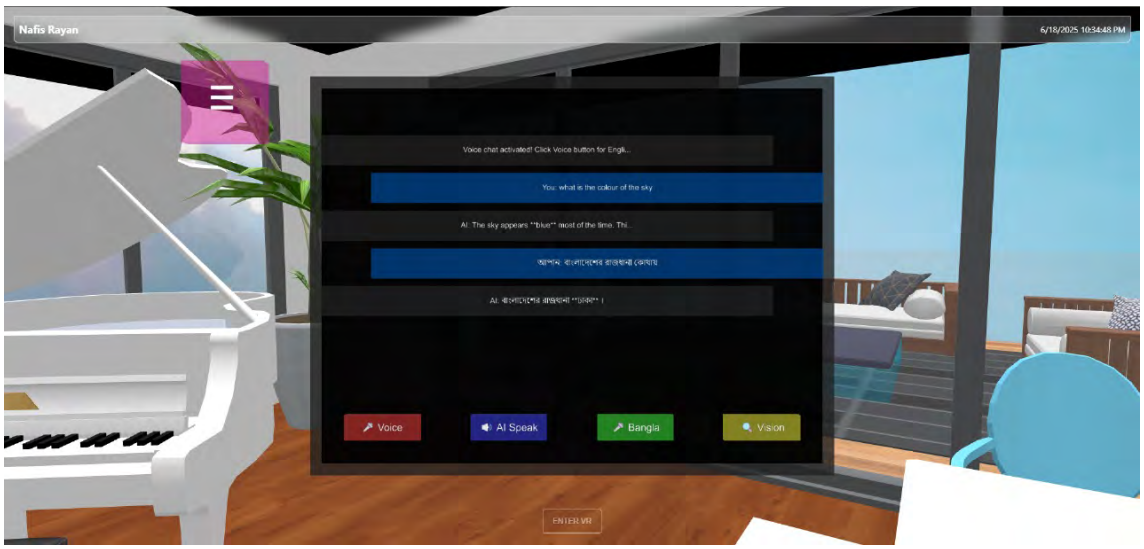

Figure 6.6: Multi-Modal Input Handling

We have developed a multi-modal system where the natural language processing (NLP) module acts as an interface between the computer vision (CV) model and the large language model (LLM). The system parses object classes and their individual coordinates from the CV model, composes them into structured, interpretable

representations, and passes them through the NLP module to facilitate seamless integration with the LLM. The structure provides uniform, context-sensitive reasoning between visual and textual domains.

This integration of the multiple modes assures that VRIA must have a reasonable notion regarding the intentions of the user and consideration of the environment virtually and in reality, thereby supporting major recommendations of system design revealed earlier.

6.5 System Architecture

All these components are wrapped up in the workable system of VRIA architecture, which allows efficient and easy data flow and processing. Advanced decision-making algorithms process inputs from LLM and CV modules to generate responses on appropriate stimuli. In addition, such architecture makes VRIA both responsive and concrete for scalability and further feature development. This system is a fine illustration of wholesome engagement with the user by ensuring smooth interaction between the LLM's linguistic capability and the environmental awareness of CV.

VRIA is an integrated system that incorporates many technologies to facilitate smooth user interaction. An integrated Bengali language LLM, Gemma2 9B, contextualizes the textual and auditory input to ensure factually grounded responses. A CV model based on YOLOv11, trained from contextually relevant datasets, provides real-time environmental perception to enhance user interaction. This, combined with auditory input from speech recognition, is processed through a multi-modal input handling system, feeding the LLM data to generate responses. Operating in a dynamic 3D environment powered by ThreeJS, VRIA permits the interaction of both physical and virtual elements, creating a responsive and flexible platform.

Chapter 7

Evaluation and Results

This chapter shows the evaluation of the Virtual Reality Intelligent Assistant (VRIA) system which is conducted by Brac University students in a Google Form. The assessment was to test the usability, performance, and effectiveness of the system, particularly for users in Bangladesh. Here, 46 undergrad students provided comments on aspects such as setup instructions, compatibility, vision and chat functionality, virtual reality interface, feature fulfillment, overall performance, and multi-modal context understanding.

7.1 Evaluation Methodology

The study employed a Google Form questionnaire distributed to Brac University students, who evaluated VRIA based on their experience with its functionality. The survey included binary (Yes/No), rating-scale (1-10), and open-ended questions for collecting extensive feedback. Quantitative metrics included ratings for setup instruction clarity, compatibility, vision capability, responsiveness of the chat system, virtual reality interface, feature accomplishment, performance, effectiveness for Bangladeshi users, and multi-modal context understanding. Qualitative responses gave feedback on user experience and areas of improvement. Statistical analysis of data was done to come to significant conclusions.

7.2 Quantitative Analysis

In our evaluation, 46 undergrad students rated VRIA aspects on a 1-10 scale (1 = poor, 10 = excellent). Below is a statistical summary of the ratings:

Metric	Mean	Median	Standard Deviation	Range
Vision Capability	7.37	8	1.45	5–9
Chat System Responsiveness	7.83	8	1.08	6–9
Virtual Reality Interface	6.46	6	1.29	5–9
Feature Fulfillment	7.93	8	1.19	6–10
Overall Performance	7.20	7	1.20	5–9
Effectiveness for Bangladeshi Users	8.07	8	0.90	7–9
Multi-Modal Context Understanding	7.50	8	1.38	5–9

Table 7.1: Evaluation Metrics Summary

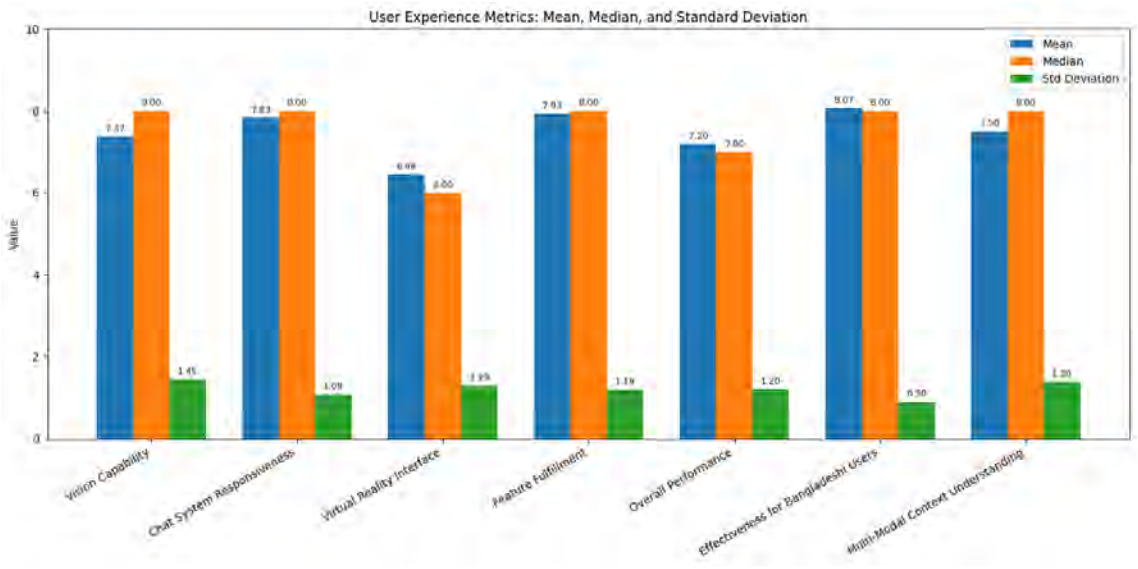

Figure 7.1: User Experience Metrics: Mean, Median, and Standard Deviation

Of 46 participants, 97.83% (45 students) have identified VRIA's setup instructions as being easy to read and adhere to, demonstrating a highly accessible setup process. Furthermore, 97.83% have also stated no problems with compatibility, although

one student experienced a problem in relation to the driver, demonstrating rare but exceptional cases requiring correction. The YOLOv11-based vision capability, which was tested using Bangladesh-specific data, was scored an average of 7.37 (SD = 1.45) with 60.87% scoring it 8 and higher, indicating value for object detection in local contexts even with the emergence of mistakes in a complex setup. The chatbot, which was created with the Gemma2 9B model that had further fine-tuning, achieved a decent mean score of 7.83 (SD = 1.08) and was praised for its response to Bengali speech and text inputs, although some mentioned problems with local dialects like Chittagonian. The Three.js-based VR interface received the lowest mean rating of 6.46 (SD = 1.29), with only 28.26% of users giving it a rating of 8 or higher, citing reasons like synchronization of audio and lack of Bengali localization as issues to be resolved substantially. Bangladeshi users' effectiveness reached an average rating of 8.07 (SD = 0.90), with 80.43% of it being rated at 8 or higher, highlighting its usage as local object, traffic, and road sign recognition, along with its educative, e-commerce, and social media applications. Multi-modal context understanding, integrating computer vision (CV), natural language processing (NLP), and large language models (LLM), scored a mean of 7.50 (SD = 1.38), with 58.70% awarding it 8 or more, while there were some suggestions to improve regional accent handling and fluent context-switching.Lastly, Feature fulfillment was rated at a mean of 7.93 (SD = 1.19), with 71.74% (33 students) giving scores of 8 or higher. VRIA met expectations for task automation and information retrieval, with suggestions for additional features like voice commands and plugin support.

7.3 Qualitative Analysis

Overall, the system's performance was described as being very good with the majority of the participants scoring high. The system was seen to be very responsive in normal conditions. There were freezing events that some users experienced when working for extended periods and lag when using it under low internet connectivity.

Open-ended feedback elicited user experience and suggestions:

- **Positive Feedback:** Participants most frequently described their experience as "Interesting" or "Good." Many appreciated viewing vision recognition being applied to unique Bangladeshi topics such as furniture, traffic flow, and street signs. They also proposed applications in e-commerce, education, and social media trend analysis. Particularly positive responses included statements such as, "The best VR project to come out of Bangladesh so far" and "Could revolutionize Bangladesh's tech ecosystem."
- **Recommendations for Improvement:** Participants suggested additional documentation for power features, mobile app version, dark mode contrast improvement, more voice command features, and support for regional dialects (e.g., Chittagonian). Additional requests included offline access, optimization for rural internet speeds, multi-user collaboration features, and more descriptive error messages. Specific feedback included requests for a light mobile version, simpler setup wizards, and integration with popular Bangladeshi apps.

7.4 Findings and Implications

VRIA has great potential in Bangladesh, reflected in the high effectiveness (8.07) and feature completion (7.93) scores. The localized vision and language capability, with Bangladesh-specific data sets and Bengali language processing, stand out as strong points. However, the VR interface, with a relatively lower score of 6.46, and some compatibility issues are areas for improvement. User feedback, such as mobile optimization, offline functionality, and dialect coverage, provides clear guidance to enhance accessibility and adoption.

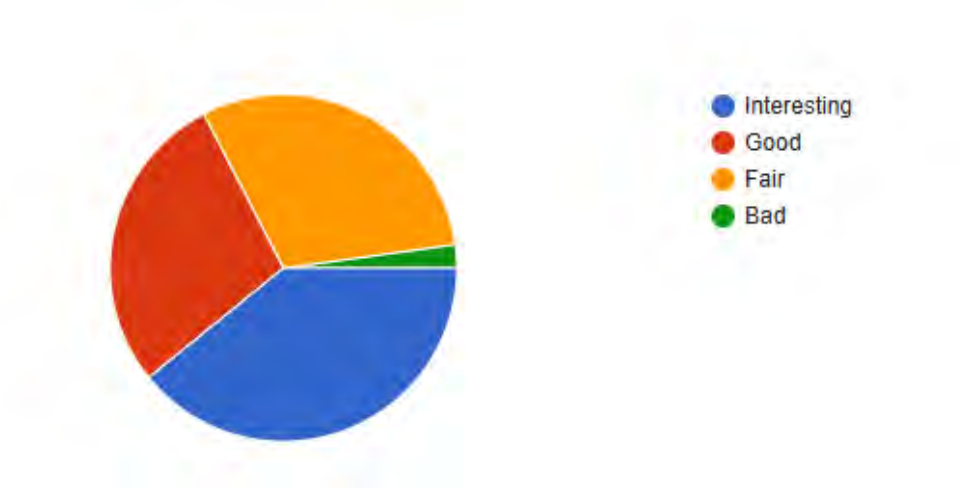

Figure 7.2: Overall Performance Pie Chart

7.5 Summary

Google Form evaluation by Brac University students, conducted from June 21 through June 22, 2025, emphasizes VRIA's merit and contextual applicability in Bangladesh, particularly in vision and chat functionality. Improving on such limitations as the VR interface and performance in a low-bandwidth environment, while accommodating user feedback, will be necessary for expanding its reach and impact. The results offer a strategic direction for subsequent development, e.g., refining the VR interface, mobile and offline optimization, and expanding dialect support, to enable VRIA to remain inclusive, effective, and impactful in diverse Bangladeshi contexts.

Chapter 8

Discussion

The Virtual Reality Intelligent Assistant (VRIA) project serves an innovative and future-facing project whose purpose is to raise the standard of the human-computer interaction inside the immersive environment through an inclusion of the multimodal technologies, specifically Computer Vision (CV), Natural Language Processing (NLP), and Large Language Models (LLMs). Able to recognise the user intent and operate contextually, VRIA is a smart, adaptive assistant that can operate in real-time object identification, gestures, and voice-based interaction, even the volume of virtual or mixed reality (VR/MR). Not only does VRIA enhance accessibility and productivity, but it also adds new opportunities to interact in resource-constrained environments such as Bangladesh by supporting multimodal inputs (visual, audio, and text) and the local language capabilities (such as Bengali). The project also preconditions the development of inclusive, smart, and scalable XR systems that fill existing gaps in technology in terms of extended reality applications.

8.1 Summary of Findings

The project has achieved its goal of creating the Virtual Reality Intelligent Assistant (VRIA), a multimodal system comprising Computer Vision (CV), Natural Language Processing (NLP), and Large Language Models (LLMs) to understand and act on user intent in immersive spaces. In real-time object detection, gesture recognition, and context-aware response generation, the system proved to be effective in a number of VR/MR applications. VRIA demonstrated to be useful in processing various types of visual, audio, and textual data, which enabled fluid human-computer interaction, improved navigation, and smart automation of virtual environments.

8.2 Contributions

The research will provide a new multi-modal paradigm that will fill the XR ecosystem gaps in the field of virtual assistant technology. It presents a modular system of connecting vision and language models adapted to real-time and context-aware interaction. The project contributes to practical innovations in intelligent UI, localised voice interaction in the Bangla language, and scalable deployment approaches with cloud and IoT technologies. VRIA opens up the prospects of an inclusive and adaptive virtual environment, particularly to the users in resource-limited settings such as Bangladesh, and improves accessibility, productivity and interaction in VR/MR applications.

8.3 Limitations

VRIA can improve engagement in extended reality, but it also deals with important issues that must be addressed. A main drawback is that these technologies require expensive devices such as Oculus, HTC Vive and HoloLens, and many people in resource-strapped places such as Bangladesh cannot afford them, so adoption is low. The virtual interface needs to be more user friendly and interactive for all users.

Even though the system runs the Bengali-adapted Gemma2 9B model, it is not always successful in interpreting many Bengali accents, dialects and how people mix Bengali and English words. Processing several input forms, like visual, auditory and textual inputs, may sometimes cause a delay or occasional inaccuracy, mostly when situations are confusing or noisy. Also, using the cloud for processing tasks makes things slower and less reliable in places with weak or unreliable internet connections. Many people are still concerned about security and privacy because visual and audio data are not made private or secure. User research, especially with different demographic groups, has not been performed, which makes it tricky to judge how people feel about using the system, their experience with it and how they handle errors.

8.4 Future Work

While VRIA could improve how people communicate in extended reality, it still must solve some major problems. A major problem is that you need advanced systems such as Oculus, HTC Vive and HoloLens, which are expensive and few people can get in countries like Bangladesh. We need to make the virtual interface more interactive and advanced for all users. In addition, while using a Bengali-based version of Gemma2 9B, the system has trouble distinguishing various regional dialects, accents and typical Bengali-English mixing elements that are very common in everyday talk. Using several senses, such as sight, sound, and text, can add to learning problems in situations with a lot of noise and when situations are complicated or confusing. At the same time, because it depends on cloud-based hardware, operation may become slow and inefficient where the internet connection isn't constant or has poor speed. Data privacy and safety are still issues because the system does not currently have encryption or anonymity processes for this data. Most importantly, many people from different demographics have not tested the system, so its ease of use, how accessible it is, and its error handling are still uncertain. Since there are still limitations, more progress and careful checks are required.

Chapter 9

Conclusion

The development of the Virtual Reality Intelligent Assistant (VRIA) is a milestone in the adaptation of modern technologies, which will boost user engagement and interaction within virtual and mixed reality environments. Through the latest advancements in Large Language Models (LLMs), Computer Vision (CV), and Natural Language Processing (NLP) technologies, alongside the deliberate integration of powerful UI, IoT devices, and Cloud Computing infrastructure, the VRIA is set to revolutionise the distinctions between the real and virtual worlds. The VRIA is a highly advanced AI-based version of a VR/MR platform, built for more effective visualisation and auditory input analysis and processing, task automation and information retrieval, personalised interaction capabilities, and much more. VRIA is set to change how we move through digital content, with augmented and virtual reality being a seamless part of our lives. Through this in-depth research, we have shown that it is viable and may bring many advantageous benefits to an adapted intelligent virtual assistant in a context of VR/MR technologies.

Bibliography

[1] G. Bouyer, P. Bourdot, and M. Ammi, "Supervision of task-oriented multimodal rendering for vr applications," in *Proceedings of the 13th Eurographics Conference on Virtual Environments*, Eurographics Association, 2007, pp. 93–100. DOI: 10.5555/2386042.2386059.

[2] P. Tiefenbacher, N. H. Lehment, and G. Rigoll, "Augmented reality evaluation: A concept utilizing virtual reality," in *Virtual, augmented and mixed reality: Designing and developing virtual and augmented environments*, Springer, 2014, pp. 226–236. DOI: 10.1007/978-3-319-07458-0_22.

[3] W. Hoenig, C. Milanes, L. Scaria, T. Phan, M. Bolas, and N. Ayanian, "Mixed reality for robotics," pp. 5382–5387, Sep. 2015. DOI: 10.1109/IROS.2015.7354138.

[4] R. Kamath and S. Natarajan, "Development of intelligent virtual environment by natural language processing," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 1, pp. 1–8, 2015. DOI: 10.14569/IJACSA.2015.060101.

[5] M. Abadi, P. Barham, J. Chen, *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, USENIX Association, 2016. [Online]. Available: https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf.

[6] T. Pfeiffer and N. Pfeiffer-Leßmann, "Virtual prototyping of mixed reality interfaces with internet of things (iot) connectivity," *i-com*, vol. 17, no. 2, pp. 179–186, 2018. DOI: 10.1515/icom-2018-0025.

[7] I. Kasyanov and A. Sirotkin, "Virtual and augmented reality in language acquisition," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 14, no. 5, pp. 4–19, 2019. DOI: 10.3991/ijet.v14i05.10018.

[8] J. R. Lykke, A. B. Olsen, P. Berman, J. A. Bærentzen, and J. R. Frisvad, "Accounting for object weight in interaction design for virtual reality," *Journal of WSCG*, vol. 27, no. 2, pp. 131–140, 2019. [Online]. Available: https://orbit.dtu.dk/en/publications/accounting-for-object-weight-in-interaction-design-for-virtual-re.

[9] M. Prabu, M. Saif, P. N. Barathi, and K. Venthan, "Evolving artificial perception by integrating nlp and cv," *International Journal of Recent Technology and Engineering*, vol. 8, no. 2, pp. 1556–1560, 2019. DOI: 10.35940/ijrte.d7867.118419.

[10] A. A. Simiscuka, T. M. Markande, and G. M. Muntean, "Real-virtual world device synchronization in a cloud-enabled social virtual reality iot network," *IEEE Access*, vol. 7, pp. 106 588–106 599, 2019. DOI: 10.1109/access.2019.2933014.

[11] X. Zhang and A. M. Smith, "Retrieving videogame moments with natural language queries," 2019. DOI: 10.1145/3337722.3341867.

[12] A. Gębczyńska-Janowicz, "Virtual reality technology in architectural education," pp. 24–28, 2020. [Online]. Available: https://mostwiedzy.pl/pl/publication/download/1/virtual-reality-technology-in-architectural-education_44451.pdf.

[13] T. Iki and A. Aizawa, "Effect of visual extensions on natural language understanding in vision-and-language models," *ArXiv.org*, Sep. 2021. DOI: 10.48550/arXiv.2104.08066.

[14] D. Lee, W. Shim, M. Lee, S. Lee, K.-D. Jung, and S. Kwon, "Performance evaluation of ground ar anchor with webxr device api," *Applied Sciences*, vol. 11, no. 17, p. 7877, 2021. DOI: 10.3390/app11177877.

[15] G. Li, D. Rempel, Y. Liu, W. Song, and C. Adamson, "Design of 3d microgestures for commands in virtual reality or augmented reality," *Applied Sciences*, vol. 11, no. 14, p. 6375, 2021. DOI: 10.3390/app11146375.

[16] F. Lu, H. Zhou, L. Guo, J. Chen, and L. Pei, "An arcore-based augmented reality campus navigation system," *Applied Sciences*, vol. 11, no. 16, p. 7515, 2021. DOI: 10.3390/app11167515.

[17] C. V. Tejaswi, "Virtual voice assistant," *International Journal for Research in Applied Science and Engineering Technology*, vol. 9, no. 1, pp. 48–52, 2021. DOI: 10.22214/ijraset.2021.35868.

[18] A. Efe, "Taking virtual reality and augmented reality to the next level," *Kamu Yönetimi ve Teknoloji Dergisi*, vol. 4, no. 2, pp. 141–165, 2022. DOI: 10.58307/kaytek.1185712.

[19] Llama, *Documentation | Llama*, Accessed: 2025-06-17, 2022. [Online]. Available: https://www.llama.com/docs/overview/.

[20] C. Meske, T. Hermanns, M. Jelonek, and A. Dogangün, "Enabling human interaction in virtual reality: An explorative overview of opportunities and limitations of current vr technology," pp. 114–131, 2022. DOI: 10.1007/978-3-031-21707-4_9.

[21] J. Wang, X. Liu, and X. Gong, "Research on intelligent methods in virtual reality games," 2022. DOI: 10.1109/cac57257.2022.10054809.

[22] R. A. Carnes and L. Casado-Mansilla, "Cloud-based nlp in virtual reality," in *Proceedings of the IADIS International Conference on Computer Science*, 2023, pp. 143–150. [Online]. Available: https://www.iadisportal.org/digital-library/cloud-based-nlp-in-virtual-reality.

[23] V. Chauhan, "Using mixed reality for iterative sketching and prototyping in product design," *International Journal for Research in Applied Science and Engineering Technology*, vol. 11, no. 6, pp. 704–742, 2023. DOI: 10.22214/ijraset.2023.53648.

[24] L. Chen, H. Zhao, C. Shi, *et al.*, "Enhancing multi-modal perception and interaction: An augmented reality visualization system for complex decision making," *Systems*, vol. 12, no. 1, p. 7, Dec. 2023. DOI: 10.3390/systems12010007.

[25] M. Chen, Z. Xing, J. Chen, C. Chen, and Q. Lu, "Enhancing virtual assistant intelligence: Precise area targeting for instance-level user intents beyond metadata," *ArXiv.org*, Jun. 2023. DOI: 10.48550/arXiv.2306.04163.

[26] S. Chen and J. Wang, "Virtual reality human–computer interactive english education experience system based on mobile terminal," *International Journal of Human-Computer Interaction*, pp. 1–10, 2023. DOI: 10.1080/10447318.2023.2190674.

[27] N. P. M. Deshpande, N. G. Chakraborty, N. T. Bhagat, N. G. Padile, and N. R. Yadav, "Intelligent companion app," *International Journal of Advanced Research in Science, Communication and Technology*, pp. 23–27, 2023. DOI: 10.48175/ijarsct-9536.

[28] C. M. Fernanda Fang, H. Huang, A. Banburski-Fahey, J. A. Fernandez, and J. Lanier, "Llmr: Real-time prompting of interactive worlds using large language models," *ArXiv (Cornell University)*, 2023. DOI: 10.48550/arxiv.2309.12276.

[29] G. Kumar, H. Dh, D. Ahmad, and H. Kiran, "Voice based virtual assistant," 2023. DOI: 10.13140/RG.2.2.32116.12163.

[30] C. Li, D. Chrysostomou, and H. Yang, "A speech-enabled virtual assistant for efficient human–robot interaction in industrial environments," *Journal of Systems and Software*, vol. 205, p. 111 818, 2023. DOI: 10.1016/j.jss.2023.111818.

[31] R. Luo, Z. Zhao, M. Yang, *et al.*, "Valley: Video assistant with large language model enhanced ability," *ArXiv (Cornell University)*, 2023. DOI: 10.48550/arxiv.2306.07207.

[32] A. Maniatis, S. Bourou, Z. Anastasakis, and K. Psychogios, "Voxreality: Immersive xr experiences combining language and vision ai models," *AHFE International*, Jan. 2023. DOI: 10.54941/ahfe1002938.

[33] P. K. Manojkumar, A. Patil, S. Shinde, S. Patra, and S. Patil, "Ai-based virtual assistant using python: A systematic review," *International Journal for Research in Applied Science and Engineering Technology*, vol. 11, no. 3, pp. 814–818, 2023. DOI: 10.22214/ijraset.2023.49519.

[34] Mistral AI, "Mistral 7b," Sep. 2023, Accessed: 2025-06-17. [Online]. Available: https://mistral.ai/news/announcing-mistral-7b.

[35] M. Mohamed and A. Sadeghi, "Integrating virtual reality and augmented reality in a collaborative user interface," *International Journal of Interactive Communication*, vol. 1, no. 1, pp. 12–19, 2023. [Online]. Available: https://ijic.utm.my/index.php/ijic/article/view/242.

[36] K. P. Rohith, H. Chinthaka, and K. Abeywardena, "Leveraging large language models for enhanced vr development: Insights and challenges," in *2023 IEEE International Conference on Virtual Reality (ICVR)*, 2023, pp. 1–6. DOI: 10.1109/ICVR57643.2023.10585495.

[37] Y. Su, "Voice2action: Language models as agent for efficient real-time interaction in virtual reality," *arXiv*, 2023. [Online]. Available: https://arxiv.org/abs/2310.00092.

[38] H. Thakkar and A. Manimaran, "Comprehensive examination of instruction-based language models: A comparative analysis of mistral-7b and llama-2-7b," in *2023 International Conference on Emerging Research in Computational Science (ICERCS)*, 2023, pp. 1–6. DOI: 10.1109/ICERCS57948.2023.10434081.

[39] K. Vaishnavi, G. P. Reddy, T. B. Reddy, N. C. S. Iyengar, and S. Shaik, "Real-time object detection using deep learning," *Journal of Advances in Mathematics and Computer Science*, vol. 38, no. 8, pp. 24–32, 2023. DOI: 10.9734/jamcs/2023/v38i81787.

[40] A. Wang, "The combination and application of virtual reality and artificial intelligence," *Applied and Computational Engineering*, vol. 6, no. 1, pp. 149–153, 2023. DOI: 10.54254/2755-2721/6/20230773.

[41] N. R. R. Yadav, N. D. Sawarkar, N. A. Dhurwade, N. P. Kawtikwar, and N. D. Pansare, "Implementing intelligent virtual assistant," *International Journal of Advanced Research in Science, Communication and Technology*, pp. 494–501, 2023. DOI: 10.48175/ijarsct-10497.

[42] s. Bande, B. Sri, D. Kadamati, *et al.*, "Revolutionizing user experience for product quality evaluation using ar/vr and nlp," *Not Provided*, vol. 1, pp. 1–6, 2024. DOI: 10.1109/icdcot61034.2024.10515452.

[43] E. Bozkir, S. Özdel, K. H. C. Lau, M. Wang, H. Gao, and E. Kasneci, "Embedding large language models into extended reality: Opportunities and challenges for inclusion, engagement, and privacy," 38:1–38:7, 2024. DOI: 10.1145/3640794.3665563. [Online]. Available: https://doi.org/10.1145/3640794.3665563.

[44] U. Community, *Yolov11: State-of-the-art object detection*, https://github.com/ultralytics/ultralytics, Open-source repository, 2024.

[45] G. DeepMind, *Gemma 2 model card*, 2024. [Online]. Available: https://huggingface.co/google/gemma-2-9b.

[46] B. Dwyer, J. Nelson, T. Hansen, *et al.*, *Roboflow (version 1.0) [software]*, https://roboflow.com, Computer vision platform, 2024.

[47] C. Farabet and T. Warkentin, *Gemma 2 is now available to researchers and developers*, 2024. [Online]. Available: https://blog.google/technology/developers/google-gemma-2/.

[48] A. Hamici, "Object detection for impaired visual assistance using transfer learning and iot-raspberry pi," University of Guelma, Tech. Rep., 2024. [Online]. Available: http://dspace.univ-guelma.dz/jspui/handle/123456789/16480.

[49] J. Izquierdo-Domenech, J. Linares-Pellicer, and I. Ferri-Molla, "Virtual reality and language models, a new frontier in learning," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 5, pp. 46–54, 2024. DOI: 10.9781/ijimai.2024.02.007.

[50] G. Jocher and J. Qiu, *Ultralytics yolo11*, version 11.0.0, 2024. [Online]. Available: https://github.com/ultralytics/ultralytics.

[51] I.-D. Juan, J. Linares-Pellicer, and I. Ferri-Molla, "Virtual reality and language models, a new frontier in learning," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, pp. 46–46, 2024. DOI: 10.9781/ijimai.2024.02.007.

[52] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements," *arXiv preprint arXiv:2410.17725*, Oct. 2024. [Online]. Available: https://arxiv.org/abs/2410.17725.

[53] M. Konenkov, A. Lykov, D. Trinitatova, and D. Tsetserukou, "Vr-gpt: Visual language model for intelligent virtual reality applications," *ArXiv.org*, May 2024. DOI: 10.48550/arXiv.2405.11537.

[54] B. Large, *Bangla-alpaca-orca*, HuggingFace, Accessed: 2024-10-20, 2024. [Online]. Available: https://huggingface.co/datasets/BanglaLLM/bangla-alpaca-orca.

[55] H. Lu, W. Liu, B. Zhang, *et al.*, *Deepseek-vl: Towards real-world vision-language understanding*, 2024. arXiv: 2403.05525 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2403.05525.

[56] Microsoft, "microsoft/Phi-3-mini-4k-instruct · Hugging Face," Jul. 2024, Accessed: 2025-06-17. [Online]. Available: https://huggingface.co/microsoft/Phi-3-mini-4k-instruct.

[57] Microsoft, *Microsoft/phi-3-vision-128k-instruct*, https://huggingface.co/microsoft/Phi-3-vision-128k-instruct, 2024.

[58] Ultralytics, *Ultralytics yolo models documentation*, Includes information on YOLOv3 to YOLO11 models, features, usage examples, and contributing guidelines. Accessed: 2024-10-21, 2024. [Online]. Available: https://docs.ultralytics.com/models/.

[59] J. Yao, X. Li, Q. Xie, *et al.*, *Llava-endo: A large language-and-vision assistant for gastrointestinal endoscopy*, 2024. DOI: 10.1007/s11704-024-40319-8. [Online]. Available: https://doi.org/10.1007/s11704-024-40319-8.

[60] R. Yousri, Z. Essam, Y. Kareem, Y. Sherief, S. Gamil, and S. Safwat, "Illusionx: An llm-powered mixed reality personal companion," *ArXiv.org*, Feb. 2024. DOI: 10.48550/arXiv.2402.07924.

[61] A. Zhalgasbayev, A. Khauazkhan, and Z. Sarsenova, "Fine-tuning the gemma model for kaggle assistant," in *2024 IEEE AITU: Digital Generation*, 2024, pp. 104–109. DOI: 10.1109/IEEECONF61558.2024.10585529.

[62] M. Abdin, S. Agarwal, A. Awadallah, *et al.*, "Phi-4-reasoning technical report," Microsoft Research, Technical Report MSR-TR-2025-05, Apr. 2025, Available as PDF from Microsoft Research website. [Online]. Available: https://www.microsoft.com/en-us/research/wp-content/uploads/2025/04/phi_4_reasoning.pdf.

[63] U. AI, *Fine-tuning guide*, Unsloth AI documentation, Accessed: 2025-06-19, 2025. [Online]. Available: https://docs.unsloth.ai/get-started/fine-tuning-guide.

[64] S. A. Chat, *Alpaca cleaned bengali*, HuggingFace, Accessed: 2025-06-16, 2025. [Online]. Available: https://huggingface.co/datasets/iamshnoo/alpaca-cleaned-bengali.

[65] K. Debanath, S. Aich, and A. Y. Srizon, "Advancing low-resource nlp: Contextual question answering for bengali language using llama," in *2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2025, pp. 1–6. DOI: 10.1109/ECCE64574.2025.11013841.

[66] Hugging Face, *Hugging Face – Blog*, Accessed: 2025-06-17, Jan. 2025. [Online]. Available: https://huggingface.co/blog.

[67] Microsoft, "microsoft/Phi-3-medium-4k-instruct · Hugging Face," May 2025, Accessed: 2025-06-17. [Online]. Available: https://huggingface.co/microsoft/Phi-3-medium-4k-instruct.

[68] Microsoft, "Welcome to the New Phi-4 Models – Microsoft Phi-4 Mini & Phi-4 Multimodal," 2025, Accessed: 2025-06-17. [Online]. Available: https://techcommunity.microsoft.com/blog/educatordeveloperblog/welcome-to-the-new-phi-4-models---microsoft-phi-4-mini--phi-4-multimodal/4386037.

[69] Qwen, *Qwen/Qwen2.5-0.5B · Hugging Face*, Accessed: 2025-06-17, Feb. 2025. [Online]. Available: https://huggingface.co/Qwen/Qwen2.5-0.5B.

[70] L. M. Schulze Buschoff, E. Akata, M. Bethge, and E. Schulz, "Visual cognition in multimodal large language models," *Nature Machine Intelligence*, vol. 7, no. 1, pp. 96–106, 2025. DOI: 10.1038/s42256-024-00963-y. [Online]. Available: https://doi.org/10.1038/s42256-024-00963-y.

[71] Ultralytics, "Yolov10," n.d. Accessed: 2025-06-17. [Online]. Available: https://docs.ultralytics.com/models/yolov10/.