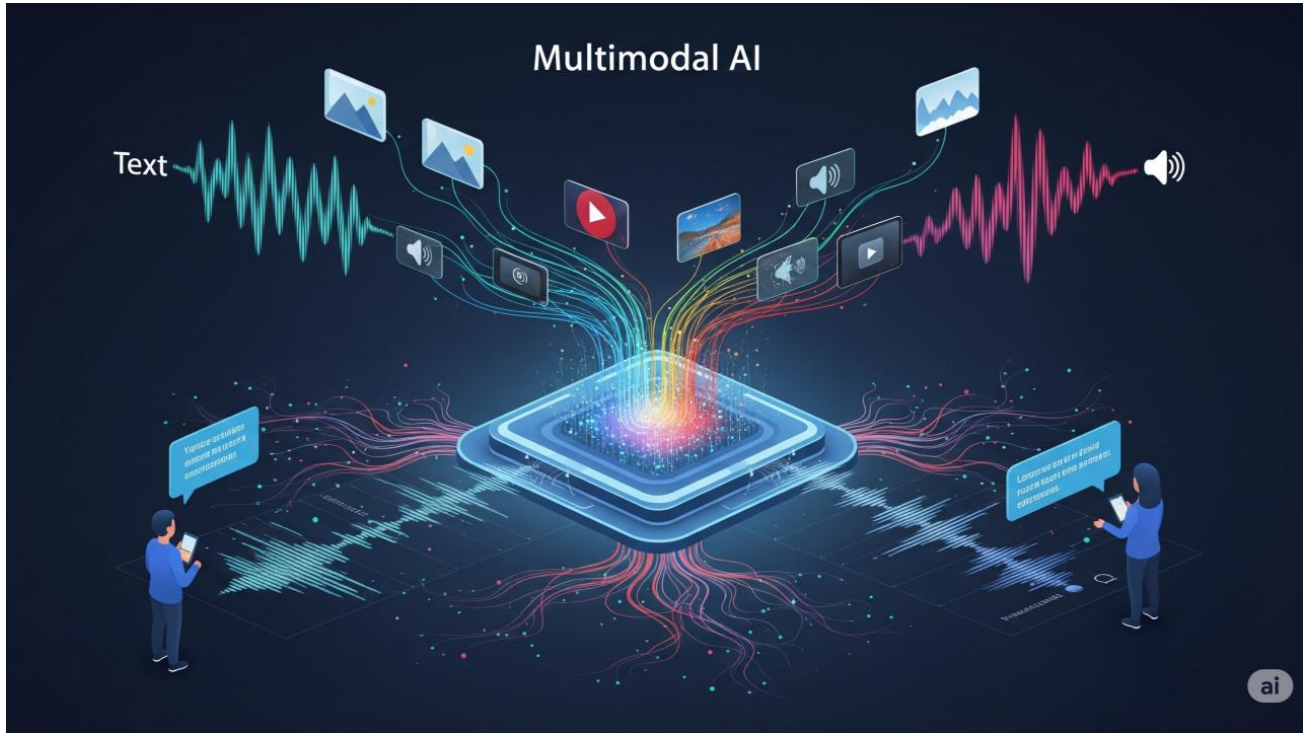
Design of a Multimodal AI System

This presentation focuses on the design of a comprehensive multimodal AI system that seamlessly integrates both vision and text modalities. It will describe the system's architecture, key components, and essential functionalities, illustrating how the blending of visual and textual data can significantly enhance understanding and performance across various applications



Introduction to Multimodal AI

Definition

Multimodal AI utilizes multiple forms of data, such as text and images, to enhance understanding.

Importance

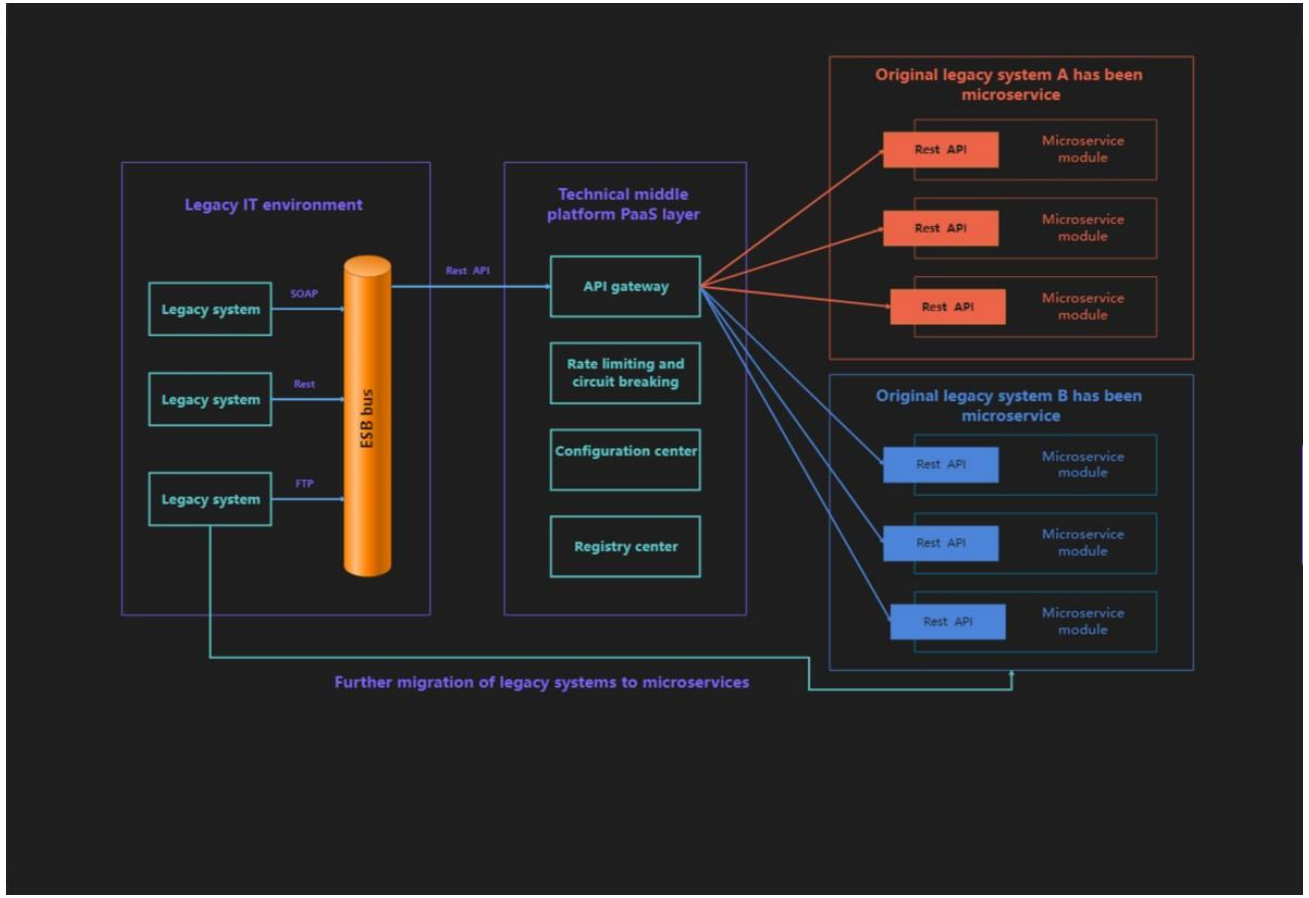
Combining modalities captures richer information, improving comprehension and decision-making.

Applications

Multimodal AI is applied in healthcare, customer service, and autonomous vehicles.

AI System Architecture

An AI system architecture is the end-to-end structural design of an AI system, detailing how components like data pipelines, machine learning models, and hardware interact to achieve business goals.



Modality Integration Techniques

Early Fusion

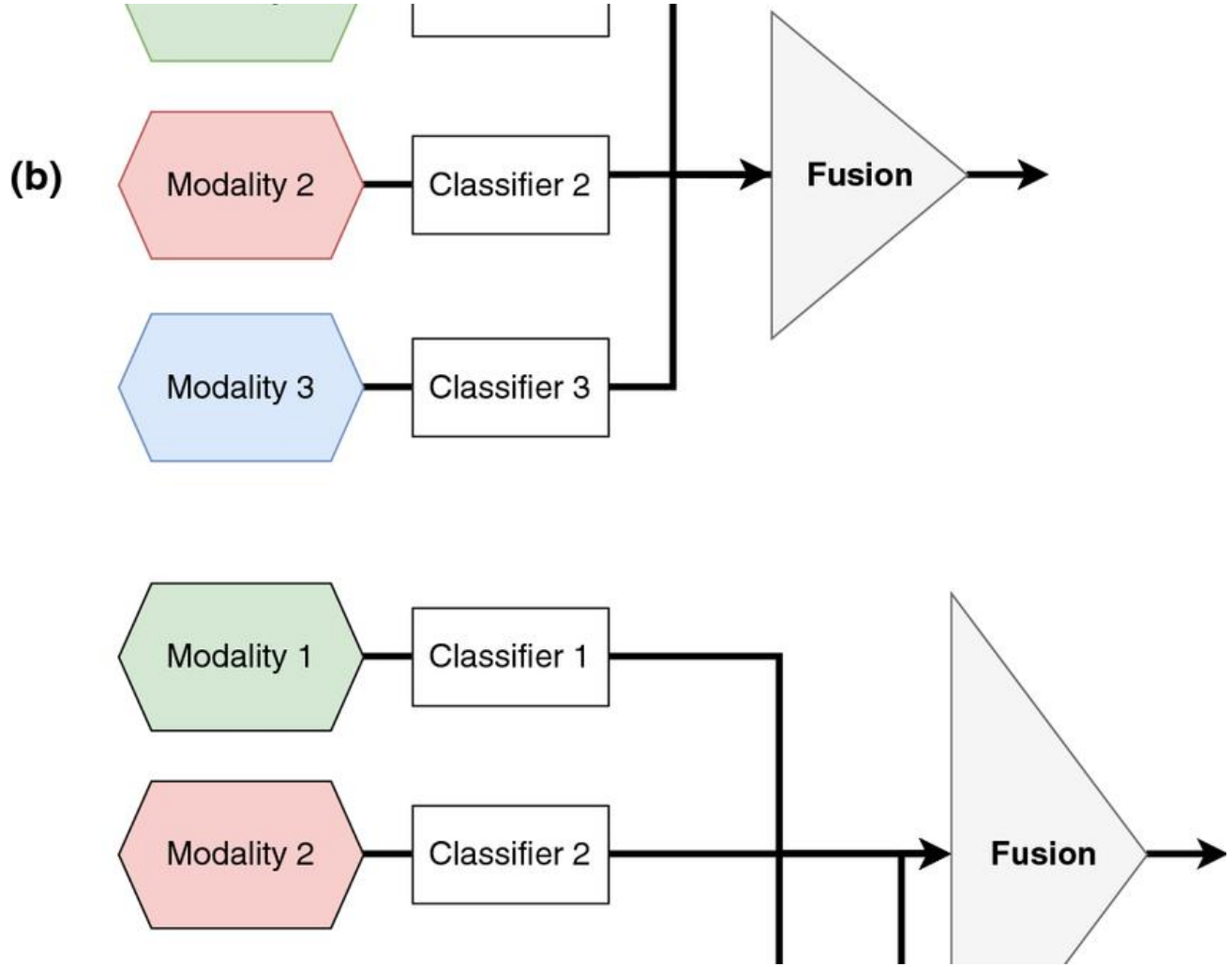
This method combines raw data from various modalities at the input stage.

Late Fusion

Here, processes for each modality run independently before their outputs are combined.

Hybrid Fusion

Combining both early and late fusion techniques, maximizing the advantages of each.



Vision and Text Modality

Vision Component

Uses CNNs to analyze visual input, identifying objects, scenes, and actions.

Text Component

Employs NLP models like transformers to understand and generate text.

Synergistic Benefits

Integration enhances tasks like visual question answering and image captioning.

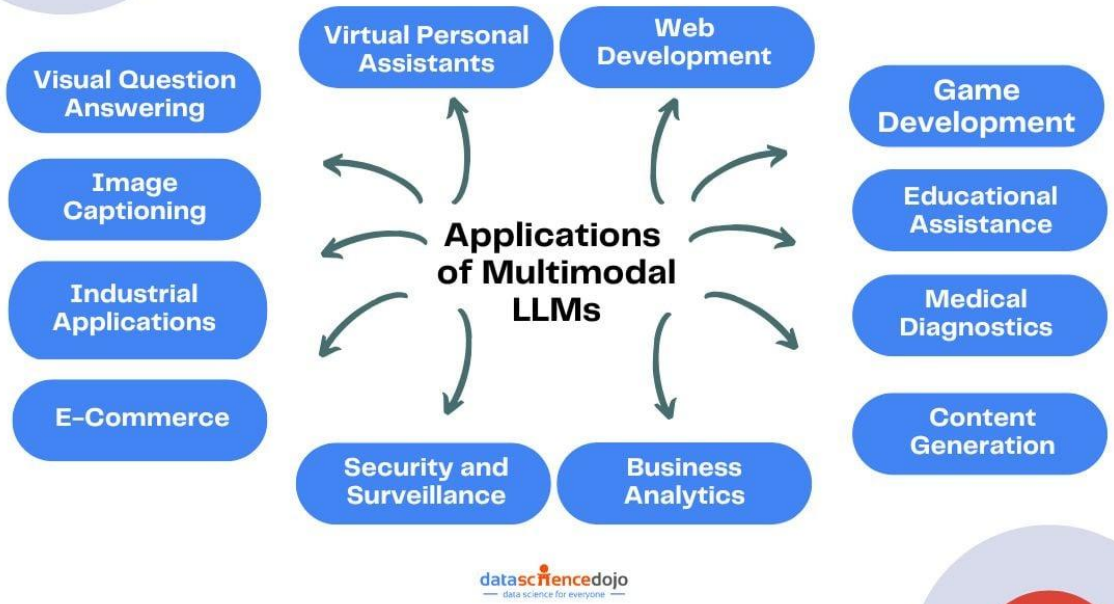
Use Case: Autonomous Driving

Autonomous Driving

refers to the use of technology and artificial intelligence to enable a vehicle to operate without direct human control. Also known as self-driving or driverless vehicles, these systems rely on a combination of sensors, cameras, radar, lidar, GPS, and onboard computing to perceive the environment, make decisions, and navigate safely from one point to another.

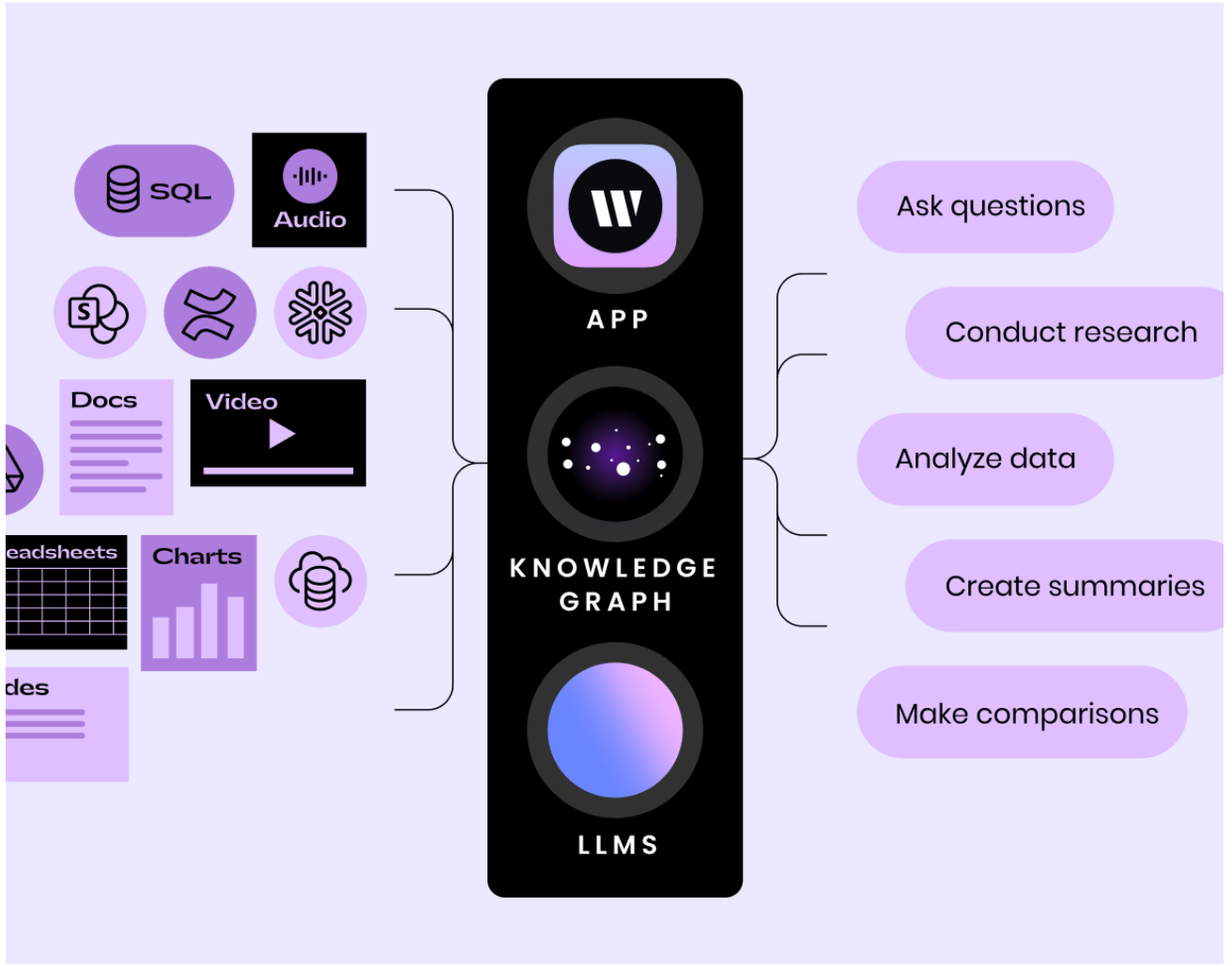
Challenges in Multimodal AI

Multimodal AI refers to artificial intelligence systems that can process and understand information from multiple modalities (e.g., text, images, audio, video, sensor data) simultaneously. It aims to replicate the human ability to integrate diverse types of sensory input to make sense of the world.



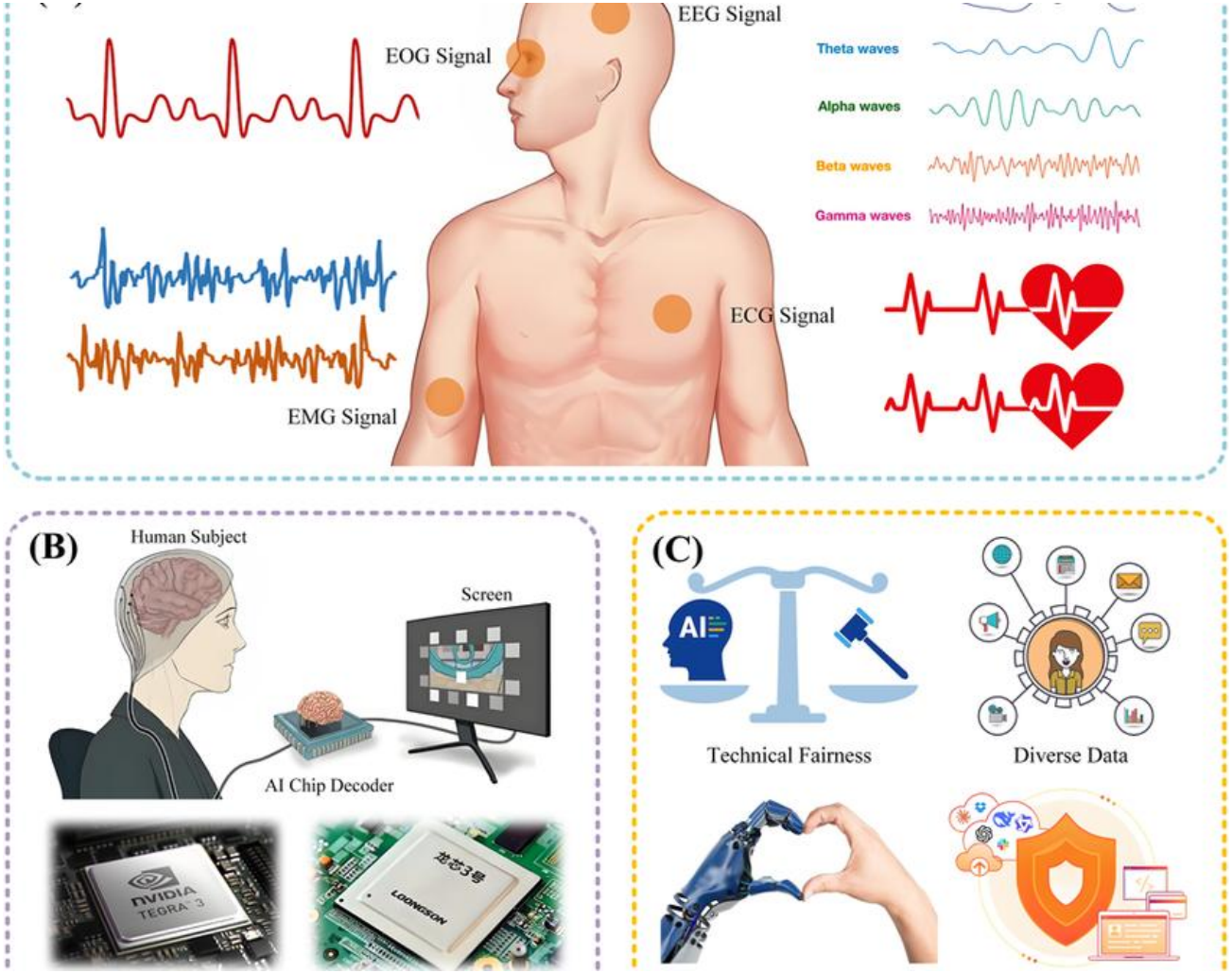
Full-Stack Multimodal AI System

A **Full-Stack Multimodal AI System** is an end-to-end artificial intelligence architecture that integrates all layers—from data acquisition and preprocessing to multimodal representation, fusion, reasoning, and user interaction—across multiple data modalities such as text, images, audio, video, and sensors



Future Directions in Multimodal AI

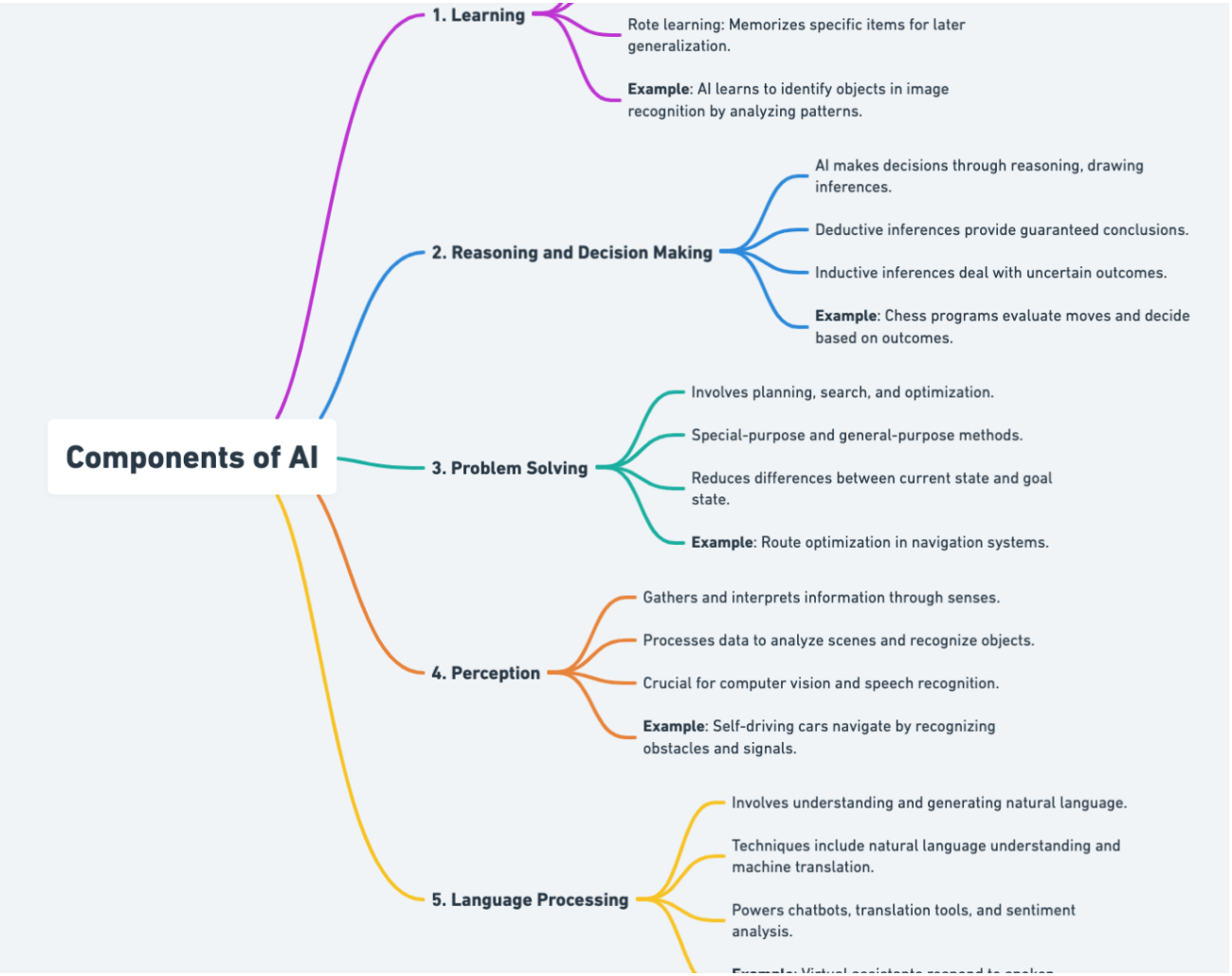
Future directions in multimodal AI refer to the emerging research trends, technological advancements, and developmental goals aimed at improving how AI systems process, integrate, and reason over multiple types of data (e.g., text, images, audio, video, sensor data) in a more intelligent, efficient, and human-like manner.



QUESTION NO 2

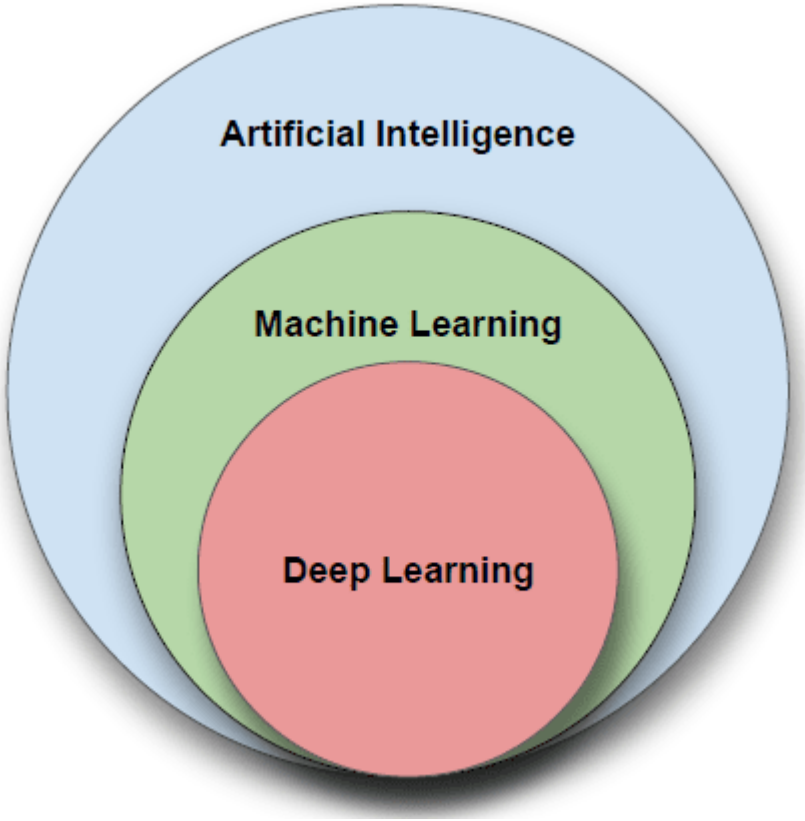
Core AI Paradigms

A **visual representation of information, concepts, or processes** that helps simplify and explain complex ideas. It usually uses shapes, lines, branches, or symbols to show how different parts are related or connected.



Overview of Artificial Intelligence

AI as a broad field encompassing various disciplines and applications, with branches like Machine Learning (ML), Natural Language Processing (NLP), Computer Vision, and Robotics

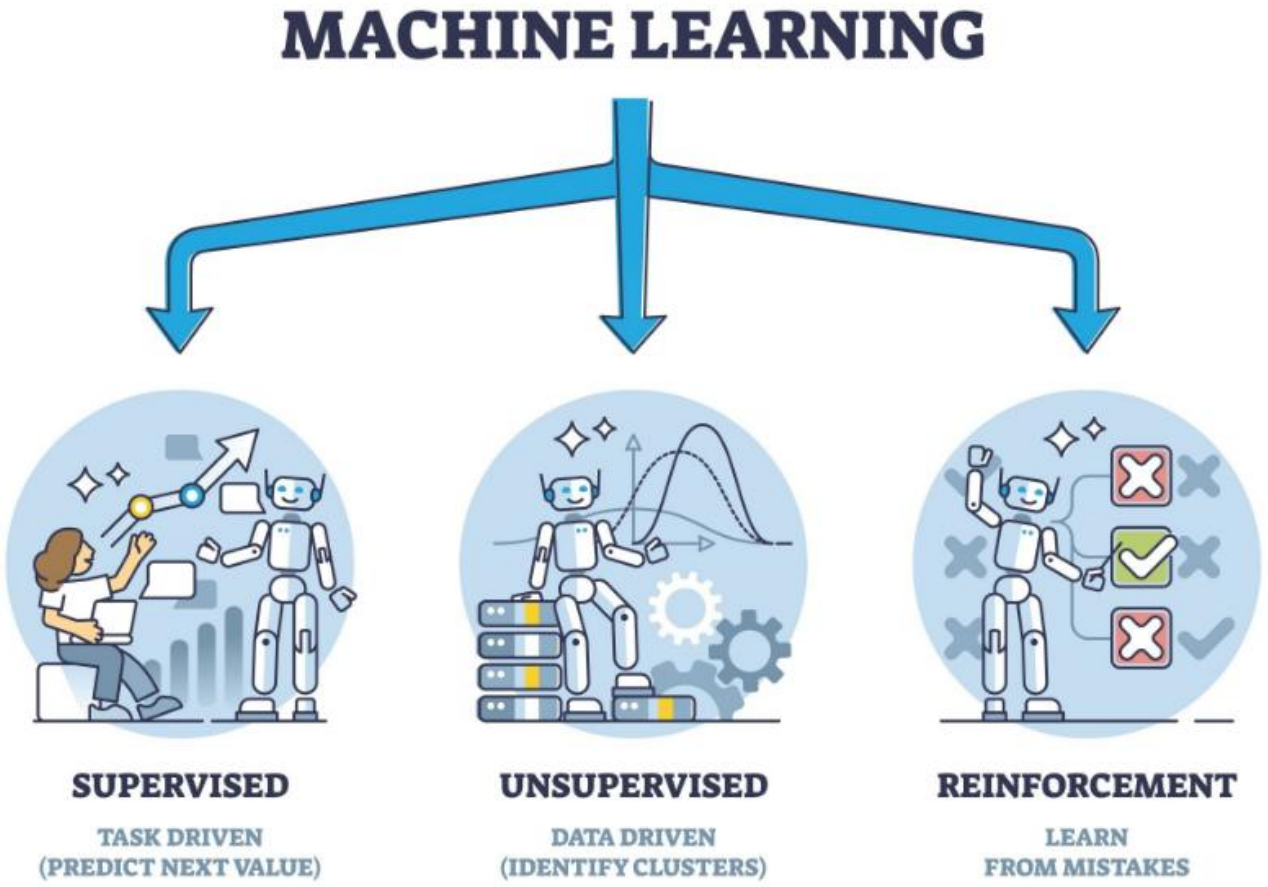


Understanding Machine Learning

Supervised Learning – Task-driven, where the system is trained with labeled data to predict outcomes (e.g., predicting house prices).

Unsupervised Learning – Data-driven, where the system identifies patterns or clusters without predefined labels (e.g., customer segmentation).

Reinforcement Learning – Trial-and-error approach, where the system learns from mistakes and rewards (e.g., training robots or game AI).



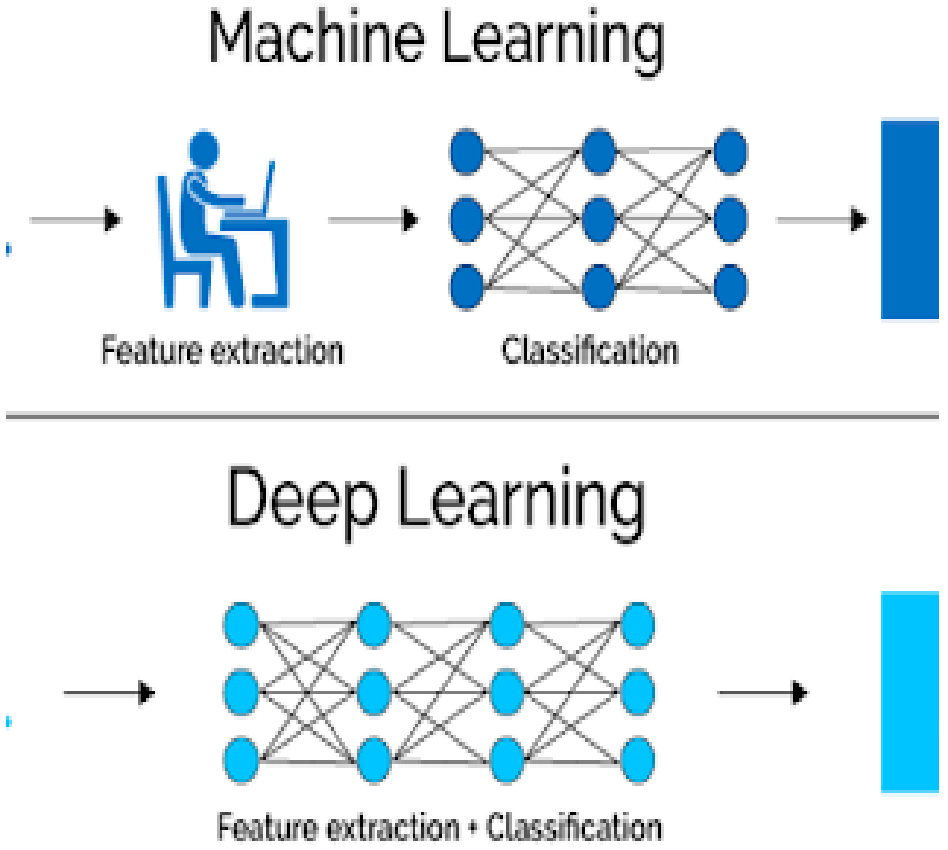
Understanding Deep Learning

Machine Learning

humans first extract features (like shapes, edges, colors) from the input data, and then the system classifies it (e.g., Car or Not Car).

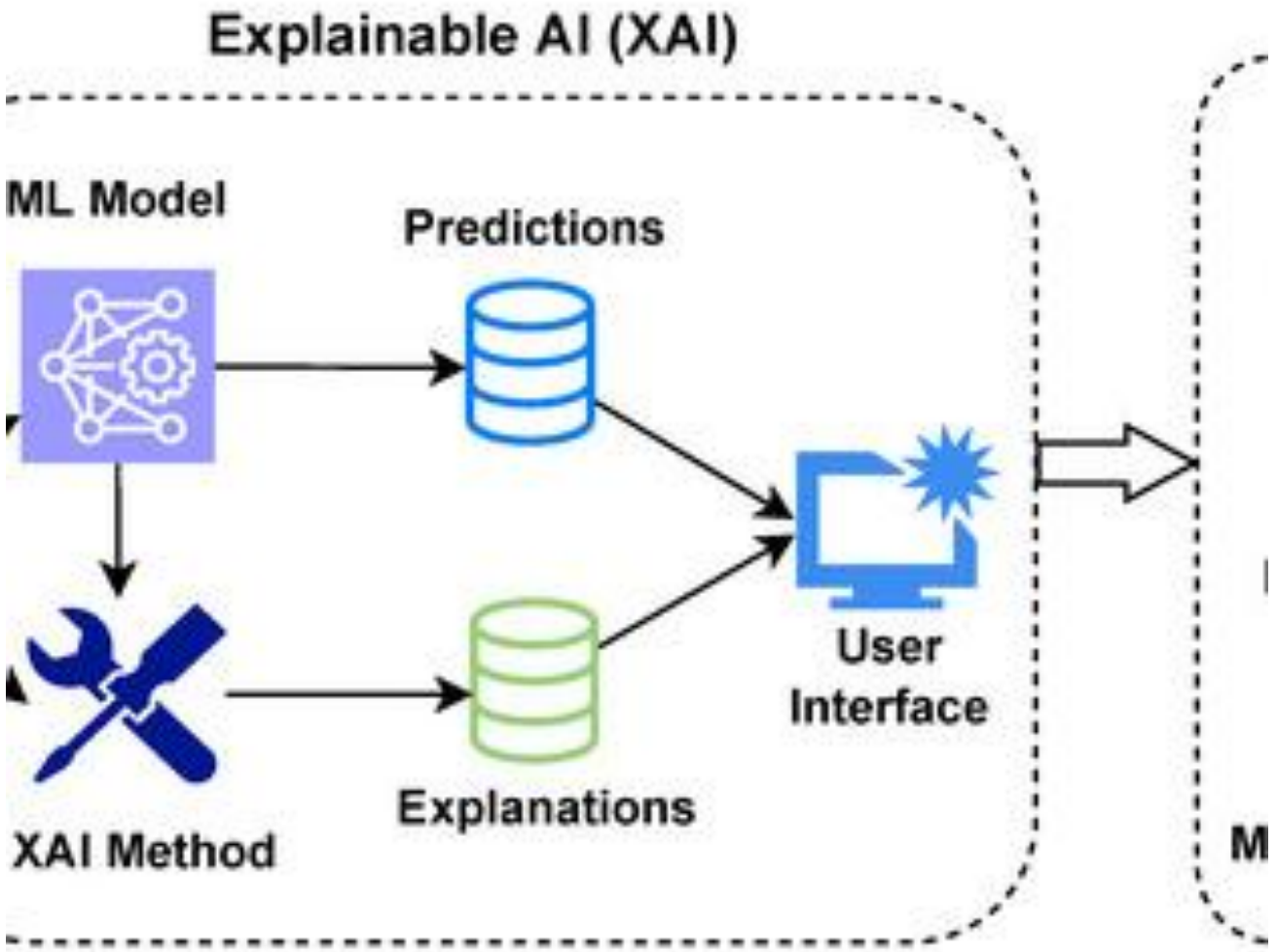
Deep Learning

he system automatically does both feature extraction and classification together without human help.



Explainable AI (XAI)

Explainable AI (XAI) takes the output of a machine learning model, adds explanations through XAI methods, and then shows both predictions and explanations to the user through an interface so the results are understandable.



Causal AI

Color-coded nodes

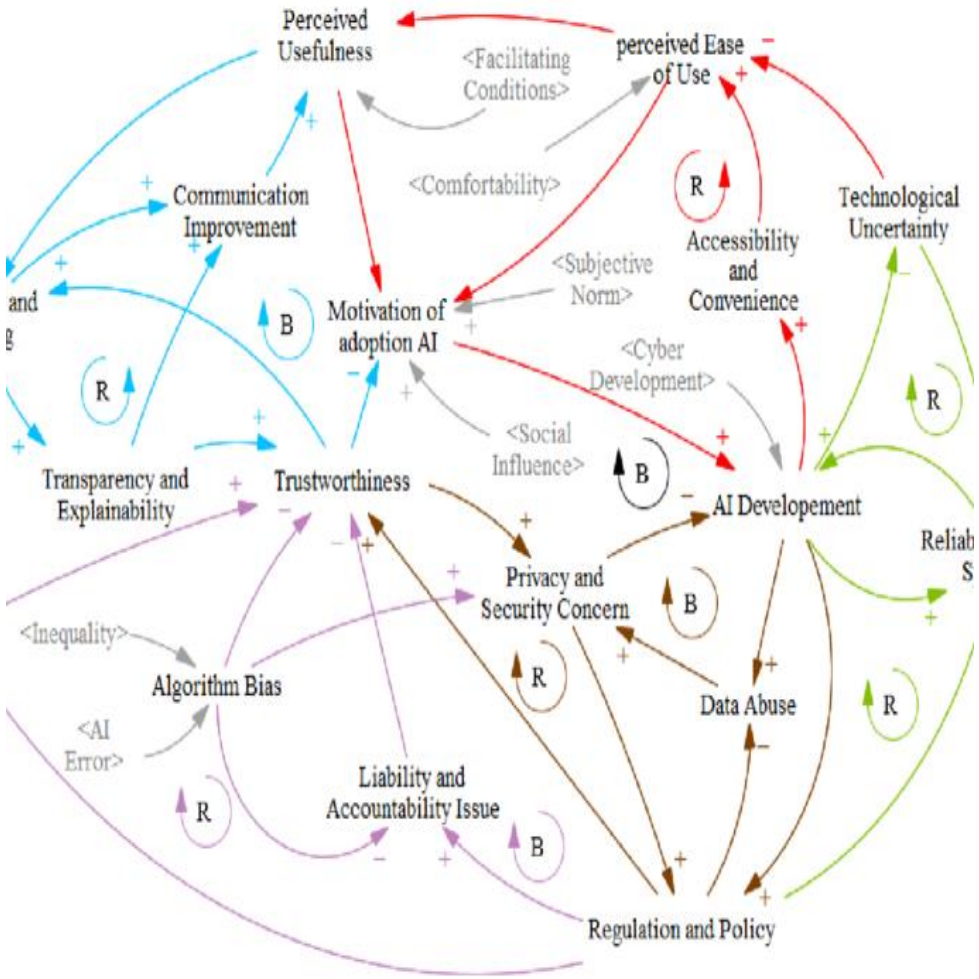
(circles/shapes) representing key concepts (e.g., "Perceived Usefulness," "Trustworthiness").

Directional arrows

showing influences or causal links between nodes.

Labels

(e.g., "R," "B") on arrows, likely denoting types of relationships (e.g., reinforcing/balancing feedback loops).



Exploring Generative AI

A hybrid AI framework integrating symbolic reasoning (explicit logic/rules) and neural networks (data-driven learning) within a knowledge graph structure to enable machines that combine human-like reasoning with adaptive pattern recognition.

Diagram of AI Paradigms

A hybrid AI framework integrating symbolic reasoning (explicit logic/rules) and neural networks (data-driven learning) within a unified knowledge graph structure.