AI Model Architecture: Modular Design Patterns

Summary

This document explores modular design in AI architectures, focusing on how models can be built as independent components that communicate via APIs. Modularization enables scalability, flexibility, and maintainability.

Core Concepts

• Layered Architecture – separating input preprocessing, model inference, and output formatting. • Microservice Model Deployment – running AI subsystems independently. • Component Interoperability – ensuring consistent data interfaces between models. • Scalability Patterns – using load balancers and distributed computing frameworks.