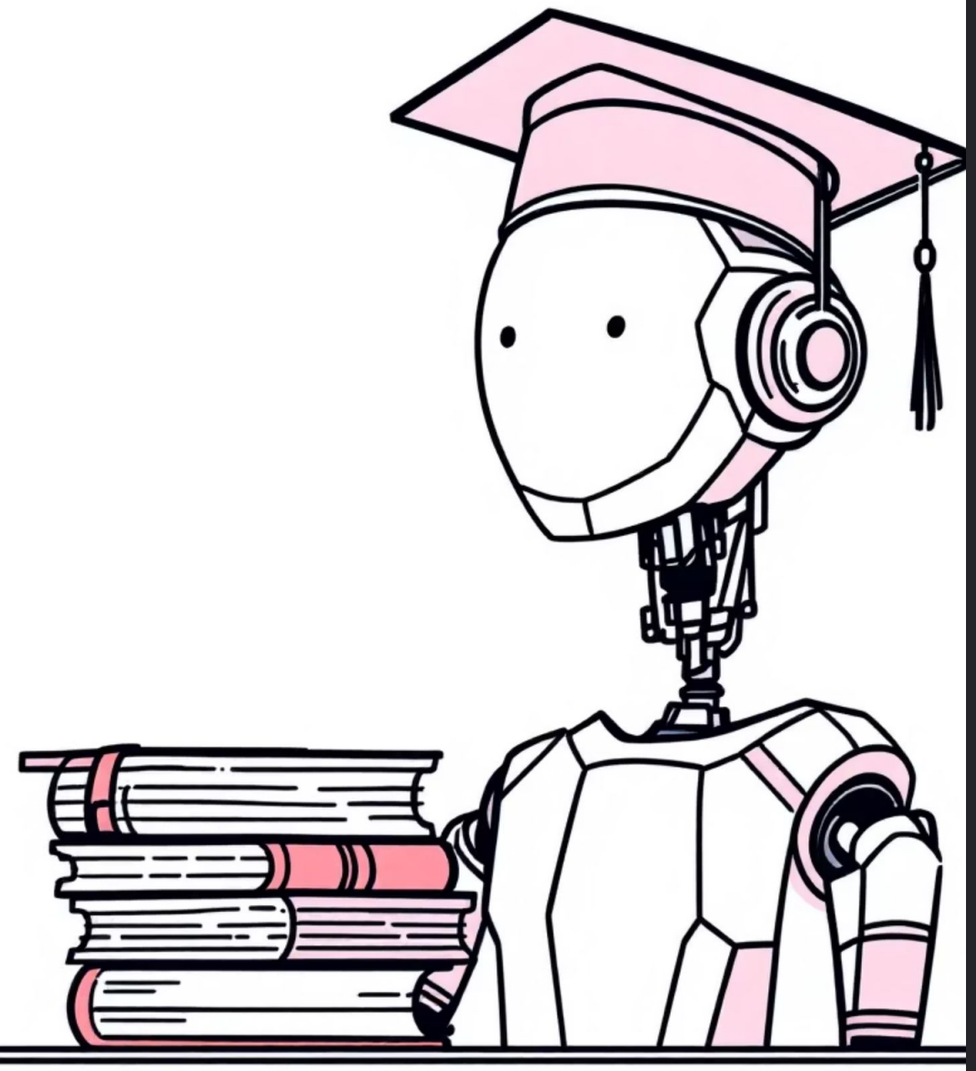


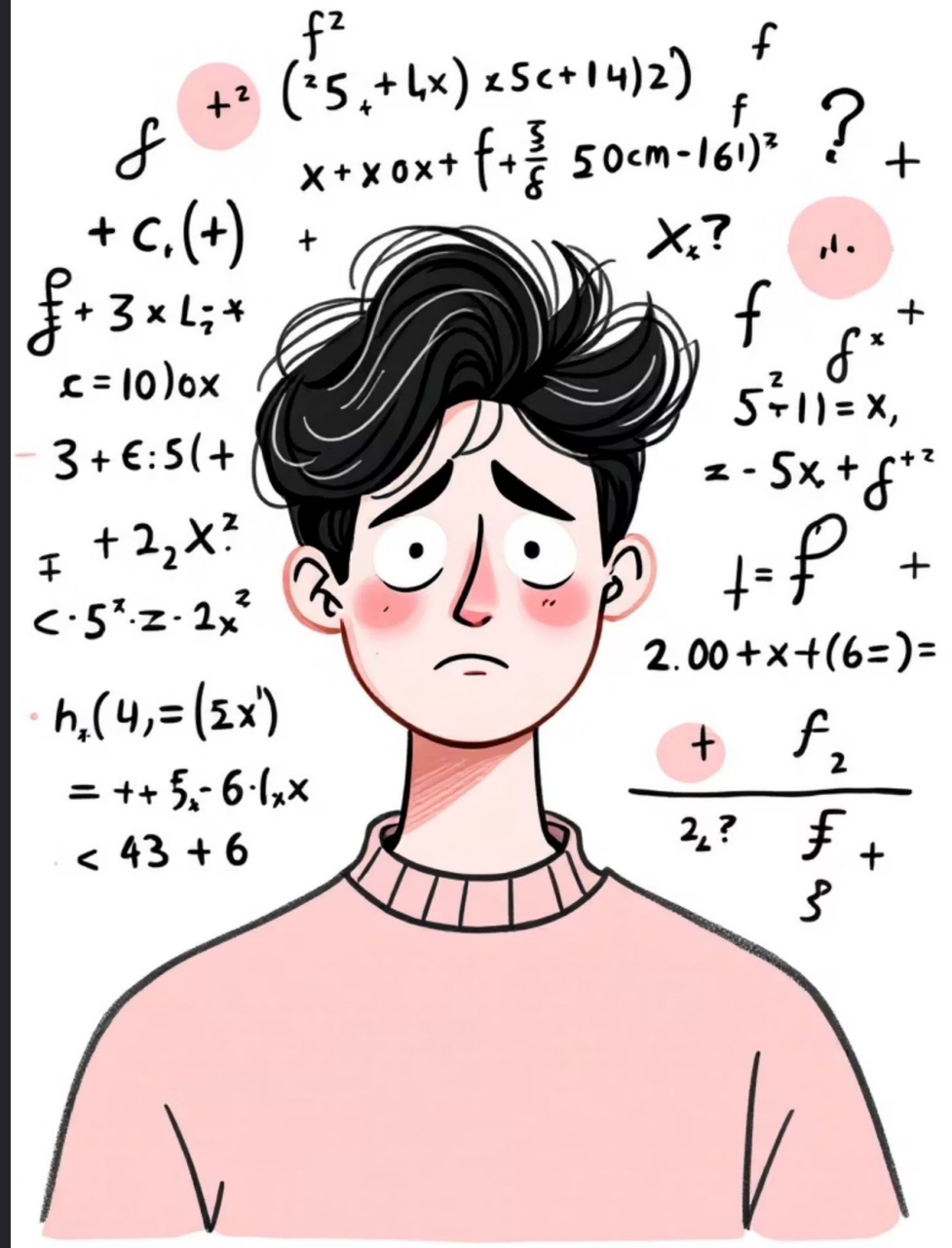
🎓 IIIT Course Assistant: RAG-Enriched LoRA Model

This project implements a sophisticated educational assistant tailored for IIIT course materials. By combining **LoRA (Low-Rank Adaptation)** fine-tuning of the **Qwen3-4B** model with a **RAG (Retrieval-Augmented Generation)** pipeline, the system provides accurate, context-aware answers to complex technical questions, further enriched with external academic resources.

Made by:
Kriti Gupta
Priyanshi Gupta

The Core Problem: Bridging the Gap

General-purpose Large Language Models (LLMs) often struggle with the nuanced, highly specific contexts of academic course materials. This project addresses the critical need for high-precision information retrieval and contextual understanding within specialized educational domains.

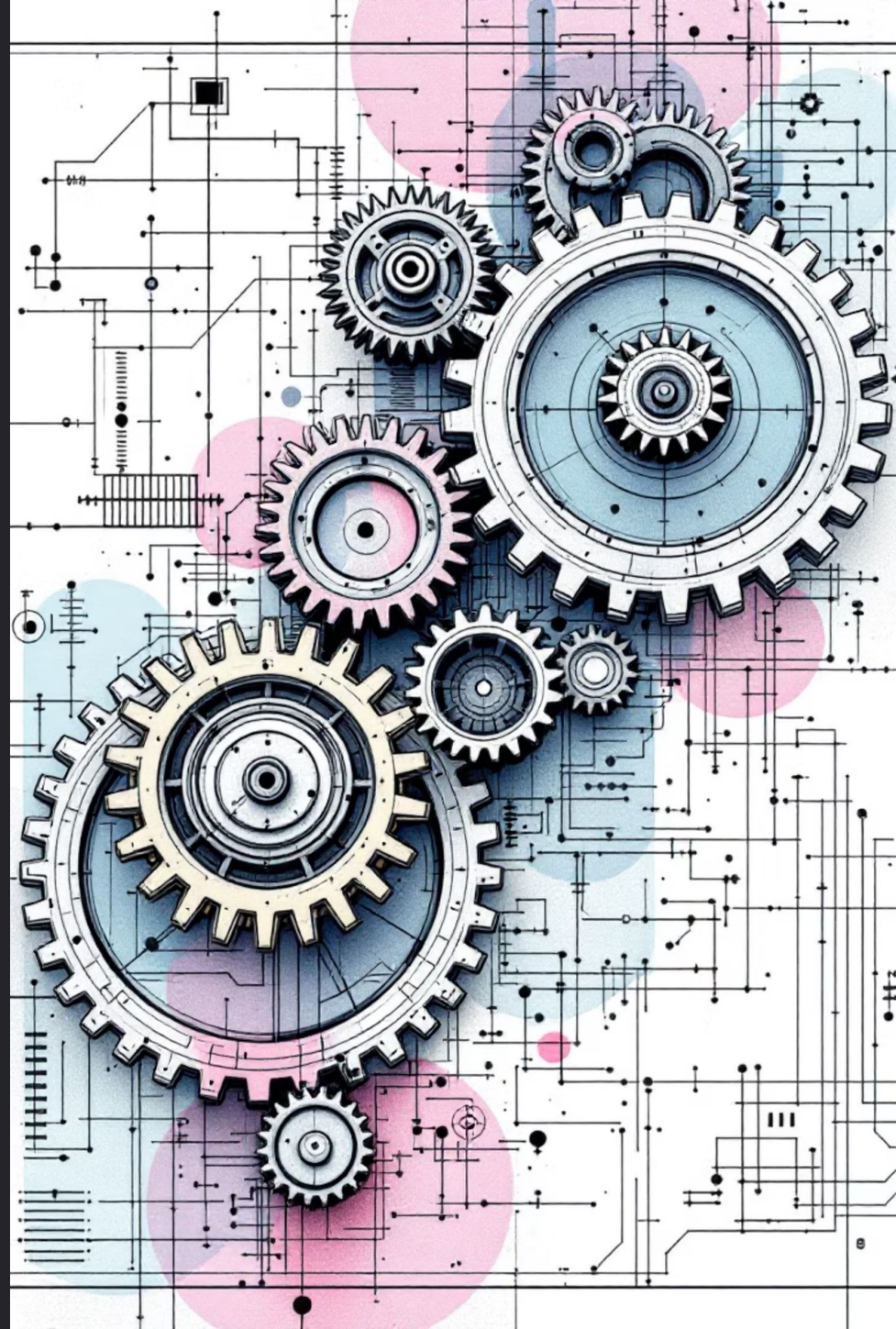


🏗️ System Architecture: A Hybrid Approach

The system employs a hybrid architecture, seamlessly blending "parametric knowledge" from fine-tuning with "external knowledge" retrieved from course documents. This modular flow ensures comprehensive and accurate responses.

🛠️ Design Choices & Justifications

The design of the **IIIT Course Assistant** is a strategic hybrid of retrieval and fine-tuning, meticulously crafted to optimize performance within the academic context.



1. Data Ingestion & Semantic Filtering

Fitz (PyMuPDF) for Extraction

Chosen for its superior handling of complex layouts and mathematical symbols common in academic PDFs, ensuring precise text extraction.

Contextual Chunking

Text is split by double newlines, targeting natural paragraph breaks that represent cohesive semantic concepts, preserving context.

The 50-Character Heuristic

Discarding chunks under 50 characters effectively filters out non-informative artifacts, enhancing the signal-to-noise ratio of retrieval.

2. Retrieval Strategy: High-Precision Baseline

The retrieval layer is engineered for exact "Source of Truth" from course materials, prioritizing accuracy over approximate methods for smaller datasets.

E5-Base-v2 Embeddings

Selected for its specialized training in **asymmetric retrieval**, excelling at matching short user queries with long technical passages.

IndexFlatL2 (Brute Force) Choice

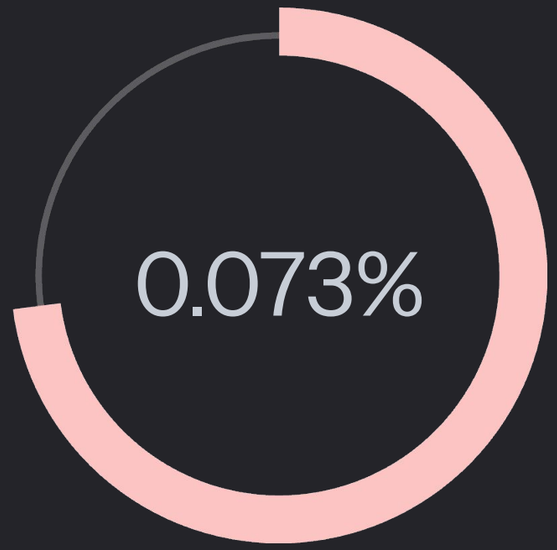
For single course data, IndexFlatL2 offers **100% accurate** retrieval with zero latency, avoiding precision loss inherent in approximate methods.

3. Model Fine-Tuning: Parameter-Efficient Adaptation

To specialize the Qwen3-4B model, a LoRA strategy was implemented, allowing efficient adaptation to academic contexts without extensive computational resources.

NF4 Quantization

Reduces VRAM usage by ~75%, enabling fine-tuning on consumer-grade GPUs like the Tesla T4 while preserving performance.



LoRA Parameter Efficiency

Only a tiny fraction (2.9 million) of parameters are trainable, focusing adaptation on attention layers (q_proj, v_proj).

Catastrophic Forgetting

LoRA adapts communication style and jargon without compromising the base model's general reasoning capabilities.

4. Enrichment Engine: Concept-Aware Tool Use

The system enhances RAG by performing **Query Transformation**, intelligently extracting concepts before engaging external APIs for broader context.

Concept Extraction

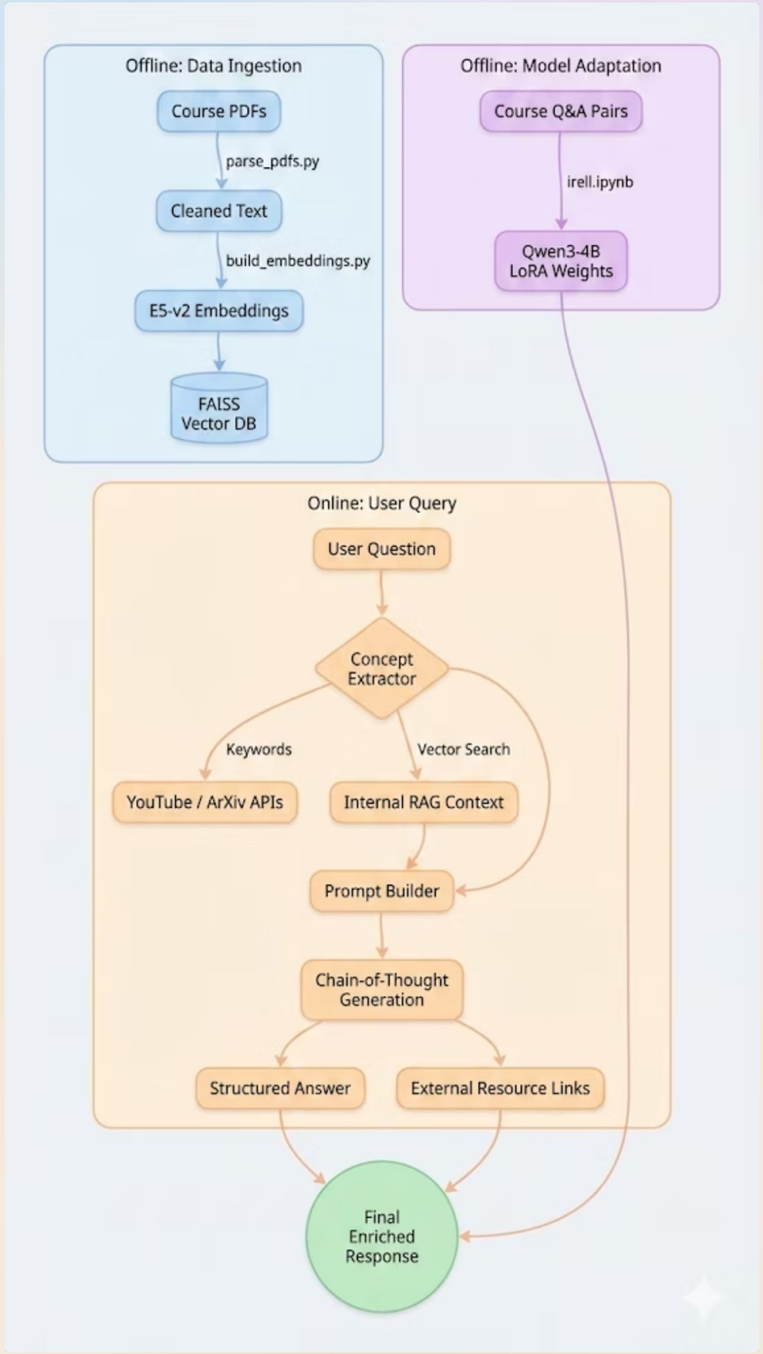
Model isolates 2-3 core mathematical/technical concepts from user queries, enabling more targeted external searches.

YouTube Search

Provides visual learners with relevant pedagogical videos, making complex topics more accessible.

ArXiv Search

Connects course content to cutting-edge research, enriching understanding with current academic developments.



🔄 System Flow: From Query to Enriched Answer

The entire process, from data ingestion to providing a synthesized, enriched response, is designed for seamless information delivery.

Feature Enrichment: Beyond Standard Question Answering

The IIIT Course Assistant transcends traditional RAG, evolving into a holistic educational tool. An intelligent Concept Extractor identifies core topics from user queries, triggering seamless integration with external academic layers for deeper understanding.

YouTube Pedagogical Search

While course materials explain 'what' to learn, YouTube offers diverse visual explanations, clarifying complex proofs and abstract concepts through varied teaching styles.

ArXiv Research Integration

Connects classroom theory to the 'latest developments' in the field. Students gain insights into cutting-edge research, bridging academic foundations with real-world applications.

Targeted Practice Retrieval

Retrieves the top 3 similar past-exam questions via vector search for targeted revision. It enriches learning by linking current queries to historical patterns, enabling exam-standard practice.

This integrated approach saves students hours of manual research, providing a single interface that automatically surfaces high-quality, peer-reviewed, and community-validated resources perfectly mapped to their specific course material and difficulty level.

Evaluation Methodology: Validating Reasoning via the IIIT Endsem

Our rigorous evaluation centers on an end-to-end testing methodology, utilizing the official IIIT Endsem solution PDFs. This approach ensures the model's performance is validated against real-world academic challenges.

A core focus of our assessment is the model's **Chain of Thought (CoT) reasoning**. We move beyond merely checking final answers, scrutinizing the internal logic displayed in the "Assistant Thoughts" field for transparency and correctness of process. The model is meticulously evaluated on its ability to dissect complex, multi-part problems into clear, logical, and sequential steps before formulating its final response.

Raw Problem Input

An unstructured question from the IIIT Endsem, requiring multi-step derivation or complex problem-solving without explicit guidance on intermediate steps.

Example: `Derive the MLE for the parameter $\lambda$ of a Poisson distribution.`

Model's Structured Derivation

The model's response, featuring a logical breakdown, intermediate steps, and LaTeX-formatted mathematical expressions for clarity and precision, aligning with academic standards.

Evaluation focuses on the correctness and coherence of these internal steps, not just the final result.

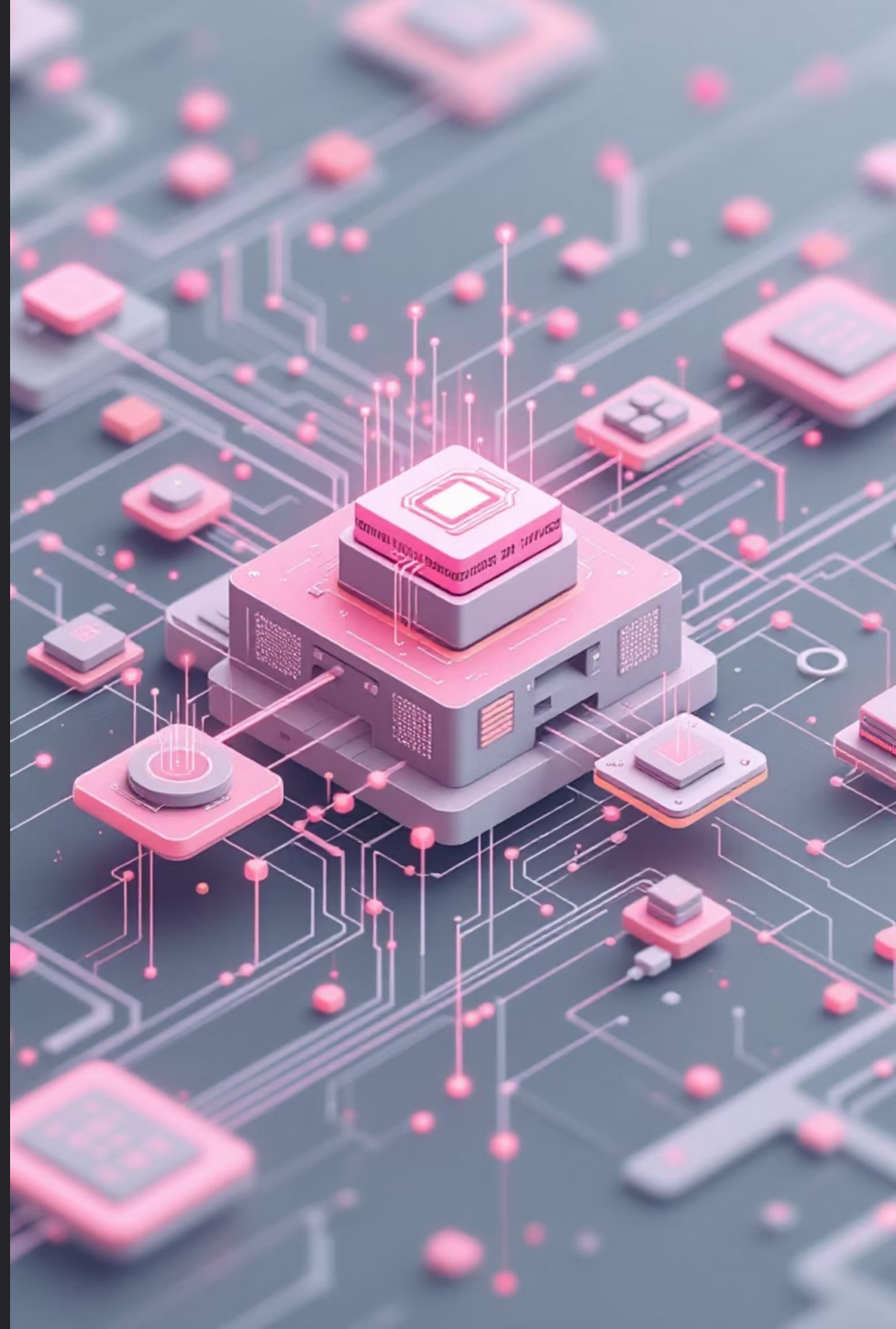
Conclusion: A Scalable Framework for IIIT Departments

Our IIIT Course Assistant is a visionary blueprint for departmental AI. Its modular architecture ensures adaptability and future-proofing across diverse academic needs.

The FAISS vector database can seamlessly transition to a centralized **Enterprise Vector DB**, while local file parsing can evolve into a **Model Context Protocol (MCP) server** for real-time document access.

Crucially, the **LoRA fine-tuning** is model-agnostic, allowing the same pipeline to scale from 0.6B to 32B models as hardware progresses.

This system automates the bridge between static course PDFs and an interactive, enriched learning ecosystem for all IIIT departments.



Developed for the IIIT iREL Task

This project successfully integrates RAG, LoRA, concept-aware external search, and automated endsem evaluation, delivering a robust and intelligent course assistant.