MavBot - UTA Course Q&A Agent

Enhanced with Course Analytics, Professor Insights & Grade Lookups

THE UNIVERSITY OF TEXAS AT ARLINGTON

Department of Data Science

Tharun Kumar Bettadalli Girish

INTRODUCTION

Course selection at UTA is difficult because information is scattered across catalogs, MavGrades, and rating sites, forcing students to manually cross-reference data.

This project creates an intelligent UTA Course Analytics & Q&A system designed to streamline how students search for and understand courses.

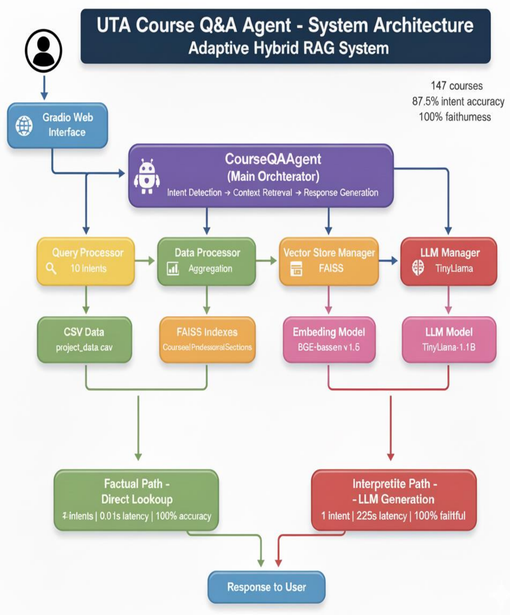
The system uses an adaptive hybrid RAG architecture, recognizing that different questions require different processing strategies.

Factual queries are answered through fast data retrieval, while interpretive or exploratory questions use vector search and LLM generation for deeper natural-language understanding.

The system interprets query intent to provide clear insights into courses, instructors, and grade trends—simplifying course discovery.

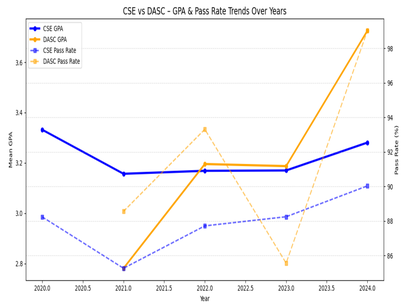
SYSTEM ARCHITECTURE

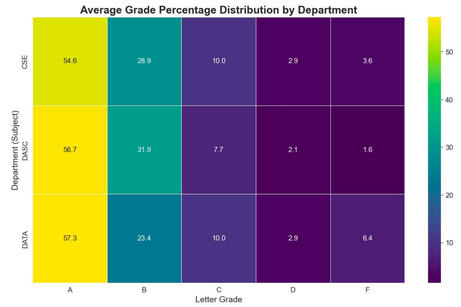
A novel hybrid architecture with intent-based routing that separates factual data retrieval from interpretive RAG processing, utilizing multi-index FAISS storage optimized for courses, professors, and sections.



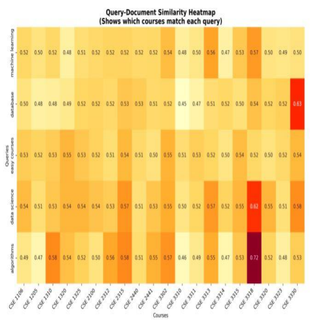
EXPLORATORY DATA ANALYSIS

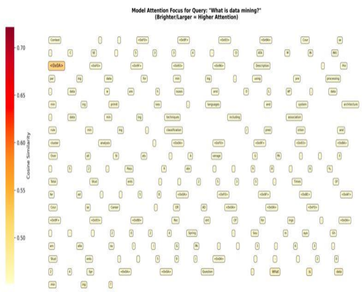
Course data from Computer Science and Data Science programs was collected via web scraping UTA catalog and MavGrades, then preprocessed through cleaning (NaN handling via department-average imputation for grade metrics and "N/A" marking for missing categorical data), standardization, feature engineering (difficulty classification, dual-chunk creation for semantic search and display), BGE-base-en-v1.5 embedding, and FAISS indexing for efficient retrieval.

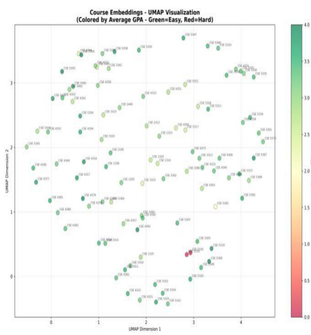


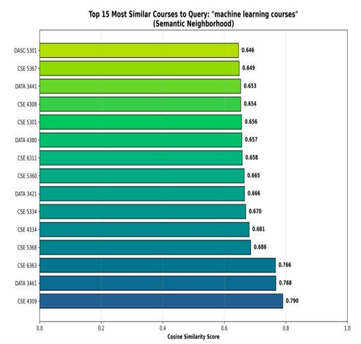


GRAPHS AND VISUALIZATIONS

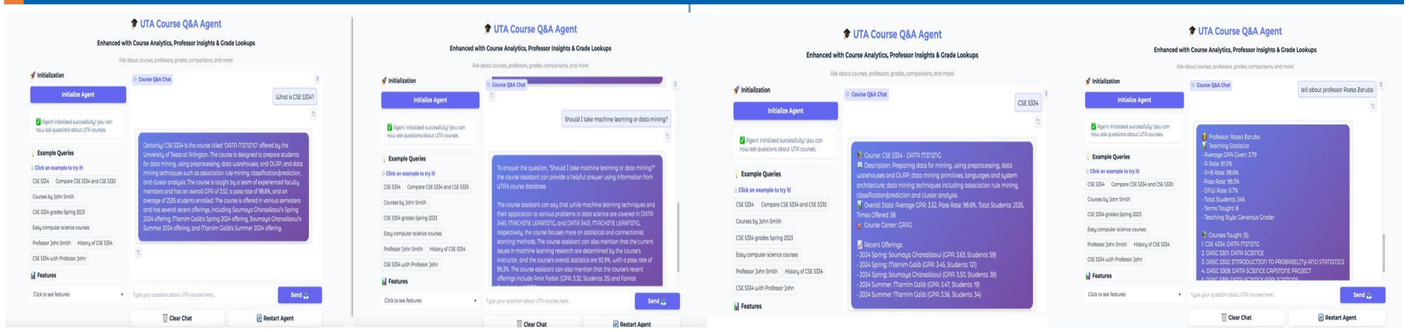








SYSTEM USER INTERFACE



SYSTEM PERFORMANCE METRICS

The factual path achieves 90% accuracy with 0.012s latency while the interpretive path maintains 100% faithfulness with natural language generation, validating the hybrid approach

Metric	Factual Path	Interpretive Path	Notes
Query Volume	90%	10%	*Based on test distribution*
Accuracy / Precision	90%	N/A	*Rule-based content checking*
Faithfulness	100%	100%	*No hallucination detected*
Relevance / Context	N/A	100%	*Context keywords present*
Average Latency	0.012s	225s	*18,750× difference*
P50 Latency	0.003s	N/A	*Median response time*
P95 Latency	0.066s	N/A	*95th percentile*
Throughput	~83 qps	~0.004 qps	*20,000× difference*

LIMITATIONS

The system faces five key constraints: CPU-based LLM latency (225s), static data requiring manual updates, regex intent detection (87.5% accuracy, fails on variations), TinyLlama-1.1B's formatting bias and limited reasoning, and scope restricted to CSE and Data Science departments

TOOLS & TECHNOLOGIES

python · BAAI · FAISS · TinyLlama

Gradio · PyTorch · pandas · NumPy

Antigravity · Google Colab · AWS EC2 · AWS SageMaker

FUTURE IMPROVEMENTS

- **GPU Acceleration**: Implement CUDA support to reduce interpretive query latency from 225s to ~5s (45× speedup)
- **ML-Based Intent Classification**: Replace regex patterns with semantic embeddings to improve accuracy from 87.5% to 95%+
- **Robust Edge Case Validation**: Add input validation and course existence checks to increase edge case handling from 33% to 90%+
- **Model Upgrade**: Transition from TinyLlama-1.1B to Phi-2 (2.7B) or Mistral-7B for improved reasoning and reduced formatting bias
- **Dynamic Data Updates**: Implement hot-reloading mechanism to refresh course data without system restart

CONCLUSION

This adaptive hybrid system proves that intelligent query routing achieves more speedup over traditional RAG for 90% of queries while maintaining 100% faithfulness, demonstrating that not all questions require expensive LLM inference. Evaluation validates production-ready performance with 90% factual accuracy (0.012s latency) and 100% LLM faithfulness, establishing a blueprint for scalable AI-powered educational systems. With clear enhancement pathways including GPU acceleration and ML-based intent classification, this work makes course discovery both instant and intelligent.