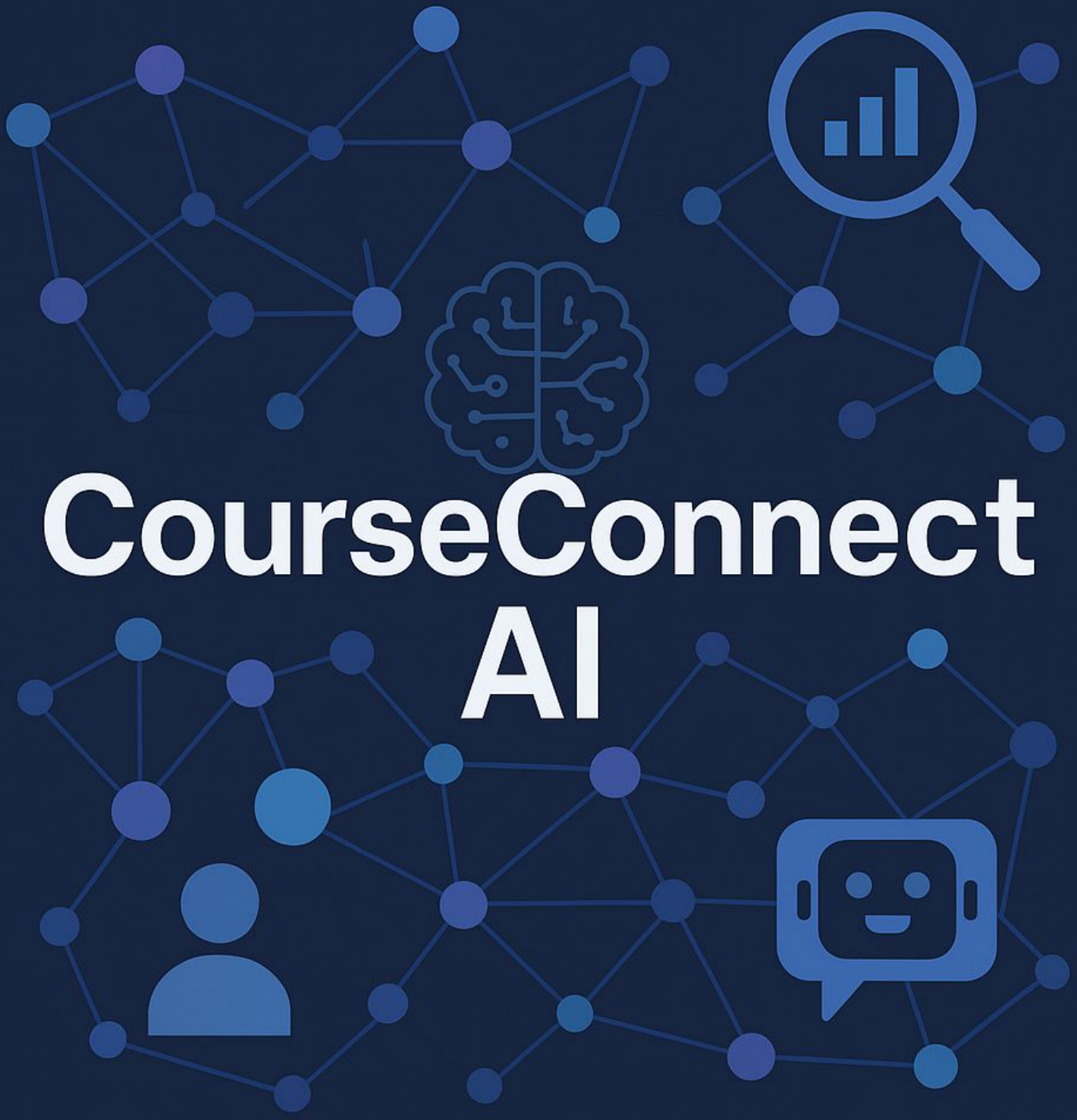

CourseConnect AI

COURSE: BUDT751-HARNESSING AI FOR BUSINESS

Team Members:

Piyali Bedagkar
Anudeep Koneru
Grace Green
James Lyons
Harsha Vardhan Simma
Sumanth Koppula
Shashank Bisht

1. Introduction

In the modern digital economy, the role of a **Data Analyst** is central to unlocking the value of data for strategic and user-focused outcomes. Analysts interpret complex data patterns, support informed decision-making, and help organizations build solutions that are both efficient and impactful.

In the context of the **Education Technology (EdTech)** sector, this role becomes increasingly critical. Institutions generate large volumes of academic data, but students often face challenges navigating course information that is scattered, outdated, or difficult to interpret. This creates a business need for solutions that improve academic planning and user engagement through data-driven design.

CourseConnect AI addresses this gap. Acting in the role of Data Analysts, our project leverages AI to transform static course catalogs into a dynamic, interactive experience making it easier for students to explore course offerings, understand relationships between topics and instructors, and plan their academic paths more effectively.

---

2. General Overview of Role / Business

As educational institutions evolve into data-rich environments, the role of the **Data Analyst** has expanded beyond reporting to becoming a strategic enabler of smarter academic services. Analysts in this space are expected to identify friction points in how students interact with course data, recommend improvements based on usage patterns, and enable personalization at scale.

In the **EdTech sector**, academic planning tools are no longer just operational utilities, they are part of the institution's digital value proposition. Students expect intuitive, intelligent platforms that mirror the experiences they have with consumer apps. This shift places data analysts at the forefront of redefining user experience within academia. Their insights shape product features, optimize course navigation, and uncover trends in course demand, enrollment timing, and instructor impact.

The demand for such analytical contributions is driven by real business goals: improving student satisfaction, boosting retention, and reducing advising overhead. Within this sector, data analysts are not just behind-the-scenes number crunchers, they are active contributors to institutional agility and student success.

---

3. Methods and Methodologies

To solve the problem of fragmented academic information, we approached the task from a **Data Analyst**'s perspective applying structured data techniques and AI-driven solutions to create a user-centric platform.

We first extracted unstructured course data from institutional PDFs and used **Natural Language Processing (NLP)** to identify key entities such as course codes, titles, instructors, schedules, and keywords. This transformed raw text into a structured format suitable for analysis.

Next, we modeled the data into a **Knowledge Graph** that visualizes connections between courses, professors, and topics enabling intuitive academic exploration. Alongside this, we integrated a **Large Language Model (LLM)** that uses this structured data to respond to natural language queries like, *"If I want a career in supply chain, which courses should I opt for?"*

Together, these methods allow **CourseConnect AI** to operate as a smart, interactive decision-support tool rooted in data analytics and student needs

---

4. Product Overview

CourseConnect AI is a smart academic planning platform built to enhance course discovery and decision-making for students.

It includes two key features:

- **Knowledge Graph**
    - Visualizes relationships between courses, instructors, and topics.
    - Helps students identify overlaps in subject areas, shared faculty, or relevant themes.
    - Encourages exploratory learning by surfacing connections not obvious in traditional catalogs.
- **Conversational Chatbot Assistant**
    - Answers course-related queries in natural language using a Large Language Model (LLM).
    - Supports questions like:
        - *"If I want a career in supply chain, which courses should I take?"*
        - *"Who teaches Python-based courses this semester?"*
    - Reduces friction in academic planning by offering real-time, personalized answers.

Together, these features:

1. Turn static course PDFs into an interactive academic tool.
2. Improve student engagement and reduce reliance on manual advising.
3. Align course choices with career goals through smart suggestions.

---

5. Technical Implementation

Data Collection & Parsing

- Extracted course data from a PDF catalog using the PyPDF2 library.
- Parsed details such as course ID, title, professor, schedule, and description into structured text using custom logic (pdf_parser.py).

Entity Extraction

- Developed an NLP-based entity extraction system using regular expressions and the **transformers** pipeline (dslim/bert-base-NER) to parse course PDFs.
- Extracted entities included:
    - Course → Title
    - Course → Professor

    - Course → Timings
    - Course → Keywords from descriptions
- Grouped these relationships into triples for graph modeling.

Knowledge Graph Construction

- Modeled extracted entities into a graph using the **NetworkX** and **PyVis** libraries.
- Converted the graph into **JSON format** for downstream integration.
- Visualized the graph in Streamlit using custom HTML/JS injection for interactivity (zoom, search, focus).
- Nodes represent courses, instructors, or topics; edges represent relationships.

LLM-Based Chatbot

- Integrated the **OpenRouter API** to connect to an **LLM (LLaMA 3 8B Instruct model)**.
- Built a prompt-driven interface that passes both course context and **graph-based JSON data** to generate relevant answers.
- Responses are generated and displayed in real-time through a conversational chatbot UI in Streamlit.

User Interface

- Developed using **Streamlit** with customized layouts and component styling.
- Features include:
    - Top navigation bar
    - Home page with product descriptions
    - Dedicated tabs for the knowledge graph and chatbot

Data Flow

- **PDF → Text → Structured Course Blocks → Entity Triples → Graph (JSON) → LLM Context**

---

6. Bias Detection and Evaluation

Source of Truth

- We used the official University of Maryland course catalog as our sole data source. As a verified institutional record, it served as the authoritative "source of truth," ensuring accuracy, representativeness, and freedom from third-party bias.
- The data was processed uniformly, minimizing inconsistencies. Course-topic-professor relationships were derived directly from structured content to ensure fair and complete coverage.

Chatbot Bias Management

- Bias was controlled by restricting the LLM's context to only university-derived data. This ensured that responses remained factual, academic, and grounded in official content—avoiding any external influence.

Model Constraints & Evaluation

- Since CourseConnect AI does not make predictions, standard bias evaluation metrics did not apply. Instead, we focused on data accuracy, structural completeness, and the contextual relevance of chatbot responses.

---

7. Future Implementation

Ideas for Improvement or Expansion

- Introduce a course recommendation system based on student interests or academic goals.
- Add user profiles to enable bookmarking, course tracking, and personalized academic planning.
- Improve chatbot capabilities with multi-turn conversations and context retention for more natural interactions.

Scalability and Long-Term Value

- Expand the system to support multiple departments or institutions beyond UMD.
- Integrate with student portals to offer tailored guidance based on enrolled programs or past courses.
- Use analytics to understand student behavior and continuously refine suggestions and graph relevance.

New Technologies to Be Adopted in the Future

- **Scalable Graph Databases**
  Migrate from basic NetworkX structures to scalable platforms like Neo4j AuraDB or Amazon Neptune for managing large academic datasets.
- **Advanced Language Models**
  Upgrade to newer LLMs such as GPT-5 or Claude 3 to enhance accuracy, contextual understanding, and multi-language support in student interactions.

---

8. Challenges

Data-Related Challenges

- The source data, extracted from institutional PDFs, was highly inconsistent and unstructured, complicating the parsing process. Additionally, course descriptions often lacked important context, limiting the AI's ability to generate complete and accurate recommendations.

Technical Hurdles

- The use of Large Language Models introduced latency, especially with complex queries. Transformer models was sensitive to formatting noise, requiring iterative refinement. Ensuring clean knowledge graph construction without duplicate or ambiguous nodes was complex.

Business/Process-Related Challenges

- Translating institutional course data into an engaging student-facing product required balancing data integrity with user experience and avoiding information overload while presenting meaningful insights through the chatbot and graph posed design trade-offs.

Mitigation Strategies

- To improve data quality, customized parsing and cleaning pipelines were developed. Continuous stakeholder engagement through pilot testing helped refine the product, ease adoption concerns, and ensure practical alignment with existing systems.

---

9. Conclusion

Summary of Findings and Impact

CourseConnect AI successfully demonstrated how structured data and AI techniques can be combined to transform static academic catalogs into interactive, user-friendly tools. By extracting key course information and modeling relationships through a knowledge graph, we enabled intuitive exploration of academic content. The integration of a domain-aware chatbot further enhanced usability, allowing students to receive immediate, relevant answers to course-related queries. This project not only addressed a clear user need in the EdTech space but also showcased how data-driven systems can improve decision-making in academic environments.

Lessons Learned and Reflections

Through this project, we learned the value of clean, consistent data pipelines and how even well-formatted sources can introduce parsing complexity. We also recognized the importance of narrowing the scope of AI models to ensure reliability and accuracy, especially in academic contexts. Most importantly, we gained hands-on experience translating analytical thinking into a tangible, user-facing product bridging technical implementation with real-world impact.

---

10. References

1. **Dong, C., Yuan, Y., Chen, K., Cheng, S., & Wen, C.** (2023). *How to Build an Adaptive AI Tutor for Any Course Using Knowledge Graph-Enhanced Retrieval-Augmented Generation (KG-RAG)*. arXiv. https://arxiv.org/abs/2311.17696
2. **Abu-Rasheed, H., Weber, C., & Fathi, M.** (2024). *Knowledge Graphs as Context Sources for LLM-Based Explanations of Learning Recommendations*. arXiv. https://arxiv.org/abs/2403.03008
3. **Jiang, W., Pardos, Z. A., & Wei, Q.** (2018). *Goal-based Course Recommendation*. arXiv. https://arxiv.org/abs/1812.10078
4. **Dawood, M.** (2023). *Assessing the Effectiveness of Chatbots in Providing Personalized Academic Advising and Support to Higher Education Students: A Narrative Literature Review*. ResearchGate. https://www.researchgate.net/publication/385304708