Lahore University of Management Sciences

AI-601 – Data Engineering for AI Systems

Instructor	Rubab Zahra Sarfraz
Room No.	TBA
Office Hours	TBA
Email	rubabsarfraz@gmail.com
Telephone	--
Secretary/TA	--
TA Office Hours	--
Course URL (if any)	https://lms.lums.edu.pk/
Support Services	LUMS offers a range of academic and other services to support students. These are mentioned below, and you are encouraged to use these in addition to in-class assistance from course staff. For a complete list of campus support services available for you click here (https://advising.lums.edu.pk/#supportservices)

Course Basics				
Credit Hours	3			
Lecture(s)	Number of Lec(s) Per Week	1	Duration	180 minutes
Recitation/Lab (per week)	Number of Lec(s) Per Week	-	Duration	-

Course Distribution	
Core	No
Elective	Yes
Open for Student Category	MS AI, SSE Undergraduates and Graduates meeting pre-requisites
Close for Student Category	--

COURSE DESCRIPTION
This course provides an end-to-end introduction to data engineering for artificial intelligence (AI) systems, covering the essential concepts and industry tools for preparing data for machine learning, including data collection, cleaning, feature engineering, validation, and model deployment. Students will gain hands-on experience with popular open-source and cloud-based data engineering frameworks to learn to address the unique data requirements and challenges in AI, such as handling large volumes, ensuring data quality, and enabling efficient, fair model training and inference. The course also focuses on leveraging Large Language Model (LLM)-based tools within data engineering workflows: students will explore LLMs for tasks like code generation, quality enhancement, and deployment automation, which streamline and enhance data engineering processes. Students will apply these skills to real-world data engineering and deployment scenarios through hands-on projects and assignments. By the end of the course, students will have developed a comprehensive understanding of data engineering principles, best practices, and practical tools tailored for AI applications.

COURSE PREREQUISITE(S)	
	AI 601: AI Augmented Software Development: - For MS AI students: AI500 Foundations of AI - For rest: CS 331 Introduction to AI or EE 514/CS 535/ CS 438 Machine Learning To take **AI Augmented Software Development**, you need foundational knowledge in programming, software development principles, and mathematical concepts like linear algebra and probability.

Tentative Grading Breakup and Policy

1. **Class Participation (5%):**   Attendance and participation in class discussions.

2. **Paper Reading (10%):**   2 during the complete semester
    a. Each student is expected to read and then write reviews about 2 selected papers.
    b. Reviews should be succinct and to the point and of maximum of 1 page in the provided template. It must cover a short summary of the paper, main contributions (technical & non-technical), strengths & weaknesses, and possible extensions?

3. **Individual Homework Assignments (30%):**   3

These hands-on homework assignments are expected to be completed individually. The classes will provide a lab-like environment for students to start on these assignments. These are designed to build the fundamental skills necessary for the students to understand how to implement most of the fundamental data engineering tasks. The assignments must be pushed to Github for grading.

4. **Group Homework Assignments (40%):** 4

These assignments are more complex as compared to individual assignments hence expected to be completed in groups. Groups should comprise 2 students, exceptions can be made for 3 students in one group. The assignments must be pushed to Github for grading.

5. **Quizzes (15%):** There will be surprise quizzes throughout the course to gauge the learning levels of students after any module. The quizzes could be paper-based or computer-based.

Laptop Policy: Students are expected to bring their laptops to classes.

Examination Details	
Midterm Exam	Yes/No: No Combine Separate: Duration: Preferred Date: Exam Specifications:
Final Exam	Yes/No: No Combine Separate: - Duration: - Exam Specifications: -

Campus supports & Key university policies

Campus Supports

Students are strongly encouraged to meet course instructors and TA's during office hours for assistance in course-content, understand the course's expectations from enrolled students, etc. Beyond the course, students are also encouraged to use a variety of other resources. (Instructors are also encouraged to refer students to these resources when needed.) These resources include Counseling and Psychological Services/CAPS (for mental health), LUMS Medical Center/LMC (for physical health), Office of Accessibility & Inclusion/ OAI (for long-term disabilities), advising staff dedicated to supporting and guiding students in each school, online resources (https://advising.lums.edu.pk/advising-resources), etc. To view all support services, their specific role as well as contact information click here (https://advising.lums.edu.pk/#supportservices).

Academic Honesty/Plagiarism

LUMS has zero tolerance for academic dishonesty. Students are responsible for upholding academic integrity. If unsure, refer to the student handbook and consult with instructors/teaching assistants. To check for plagiarism before essay submission, use similarity@lums.edu.pk. Consult the following resources: 1) Academic and Intellectual Integrity (http://surl.li/gpvwb), and 2) Understanding and Avoiding Plagiarism (http://surl.li/gpvwo).

LUMS Academic Accommodations/ Petitions policy

Long-term medical conditions are accommodated through the Office of Accessibility & Inclusion (OAI). Short-term emergencies that impact studies are either handled by the course instructor or Student Support Services (SSS). For more information, please see Missed Instrument or 'Petition' FAQs for students and faculty (https://rb.gy/8sj1h)

LUMS Sexual Harassment Policy
LUMS and this class are a harassment-free zone. No behavior that makes someone uncomfortable or negatively impacts the class or individual's potential will be tolerated.

To report sexual harassment experienced or observed in class, please contact me. For further support or to file a complaint, contact OAI at oai@lums.edu.pk or harassment@lums.edu.pk. You may choose to file an informal or formal complaint to put an end to the offending behavior. You can also call their Anti-Harassment helpline at 042-35608877 for advice or concerns. *For more information: Harassment, Bullying & Other Interpersonal Misconduct: Presentation (http://surl.li/gpvwt)*

Lahore University of Management Sciences

Course Learning Outcomes	
CLO1 **CLO2** **CLO3** **CLO4**	By the end of the course, the students should be able to: ● Design and implement robust ETL and AI-specific data pipelines, including batch and real-time processes. ● Utilize LLM-based tools for enhancing data engineering workflows through automation, code generation, and validation. ● Employ data orchestration tools to manage complex workflows and ensure fault-tolerant AI applications. ● Understand deployment strategies for AI models, including integration, monitoring, and maintenance.

Relation to Program Learning Outcomes

CLOs	Related PLOs	Level of Learning	Teaching Methods	CLO Attainment checked in
CLO1	PLO1: Academic Education	Cog-3	Instruction, Homework	Assignments
CLO2	PLO2: Knowledge for Solving Computing Problems	Cog-4	Instruction, Homework	Assignments
CLO3	PLO2: Knowledge for Solving Computing Problems	Cog-4	Instruction, Homework	Assignments
CLO4	PLO2: Knowledge for Solving Computing Problems	Cog-4	Instruction, Homework	Assignments

Textbook(s)/Supplementary Readings
Textbook: Class material **Reference material**:   o   Supplemental material including tutorial videos and demonstrations for every module will be shared before the start of every module.

Lahore University of Management Sciences

Week		Topics	Assessment
COURSE TOPICS			
1-2	Module 1: Foundations of Data Engineering for AI	• Introduction to Data Engineering for AI: Overview of the data engineering lifecycle in AI workflows and the importance of clean, validated data for AI. • Core Concepts and Tools: Introduction to data pipelines, different data formats, ETL processes, and popular frameworks.	Exercise: Building a basic ETL pipeline using core tools.
3-4	Module 2: LLMs in Data Engineering Workflows	• Introduction to LLM-Based Tools: Overview of how Large Language Models (e.g., GitHub Copilot, Cursor, Colab) support data engineering tasks like code generation, quality enhancement, and automation. • LLMs for Workflow Automation: Using LLMs to streamline code writing, documentation, and basic data validation tasks.	Exercise: Applying LLMs to automate repetitive tasks and improve data quality processes in a data pipeline.
5-6	Module 3: Data Quality and Validation in Real-World Datasets	• Ensuring Data Quality: Types of data quality issues and their impact on AI model performance. • Advanced Validation Techniques: Using LLMs and traditional methods to detect and address data quality issues, focusing on real-world imperfect datasets. • Fairness in AI: Importance of data quality in ensuring fair and trustworthy inferences.	Exercise: Real-world data quality assessment and validation, leveraging LLMs to assist in the process.
7-8	Module 4: Advanced Data Pipeline Design	• Building Complex Data Pipelines: Design and implement batch and real-time pipelines for AI applications for different data formats. • Data Transformation for AI: Feature engineering and data transformations specific to ML model requirements. • Data Versioning and Governance: Managing data versions, ensuring quality, and handling metadata collaboratively.	Exercise: Building a training and testing data pipeline for an AI application.
9-10	Module 6: Data Orchestration for AI Pipelines	• Introduction to Data Orchestration: Overview of orchestration tools like Prefect and Apache Airflow. • Workflow Management: Creating, scheduling, and monitoring pipelines for AI applications. • Error Handling and Recovery: Managing failures and ensuring fault-tolerant workflows.	Exercise: Implementing an orchestrated ETL pipeline using Prefect or Airflow.
11-12	Module 5: Deploying AI-Enhanced Applications & their Maintenance	• Integration of AI Models: Integrating pre-trained models into applications through APIs and embedding models within software. • Deployment Considerations for AI Models: This section provides steps for deploying AI models in applications, focusing on performance, security, and scalability. • DevOps for AI: Introduction to DevOps concepts tailored to AI applications, including principles of CI/CD in AI workflows. • Production Monitoring of AI Models: Techniques for monitoring models post-deployment and managing AI systems in production environments.	Exercise: Deploying a simple AI model within a web or API-based application.
13	Module 7: Case Studies, Industry	• Case Studies in AI-driven Data Engineering: Review industry use cases (e.g., Microsoft, Google, Amazon) using AI in software and data engineering workflows.	

	Applications, and Ethics	• Ethics and Responsible AI: Discuss ethical considerations, biases, and responsible AI use in data engineering. • Industry-Standard Tools and Emerging Trends: Overview of current tools and trends enhancing software quality, productivity, and responsible data engineering practices.	
14	Course Closure and Discussion	• Review of Key Concepts: Recap the major topics covered, including data engineering principles, quality assurance, cloud-based data storage, and model deployment. • Discussion on Real-World Applications: Reflect on previous week's case studies and discuss how course concepts apply to industry scenarios. • Revisiting Ethical and Responsible AI: Reflection on perspectives on ethical considerations, fairness, and responsible AI use, reinforcing the importance of bias mitigation. • Q&A and Open Forum: Allowing students to ask lingering questions, discuss challenges, or clarify any areas they want to revisit.	

Academic Honesty

The principles of truth and honesty are recognized as fundamental to a community of teachers and students. This means that all academic work will be done by the student to whom it is assigned without unauthorized aid of any kind. Plagiarism, cheating and other forms of academic dishonesty are prohibited. Any instances of academic dishonesty in this course (intentional or unintentional) will be dealt with swiftly and severely. Potential penalties include receiving a failing grade on the assignment in question or in the course overall. For further information, students should make themselves familiar with the relevant section of the LUMS student handbook.

Harassment Policy

SSE, LUMS and particularly this class, is a harassment free zone. There is absolutely zero tolerance for any behavior that is intended, or has the expected result of making anyone uncomfortable and negatively impacts the class environment, or any individual's ability to work to the best of their potential.

In case a differently-abled student requires accommodations for fully participating in the course, students are advised to contact the instructor so that they can be facilitated accordingly.

If you think that you may be a victim of harassment, or if you have observed any harassment occurring in the purview of this class, please reach out and speak to me. If you are a victim, I strongly encourage you to reach out to the Office of Accessibility and Inclusion at oai@lums.edu.pk or the sexual harassment inquiry committee at harassment@lums.edu.pk for any queries, clarifications, or advice. You may choose to file an informal or a formal complaint to put an end of offending behavior. You can find more details regarding the LUMS sexual harassment policy here.

To file a complaint, please write to harassment@lums.edu.pk.

SSE Council on Equity and Belonging

In addition to LUMS resources, SSE's **Council on Belonging and Equity** is committed to devising ways to provide a safe, inclusive and respectful learning environment for students, faculty and staff. To seek counsel related to any issues, please feel free to approach either a member of the council or email at cbe.sse@lums.edu.pk

Rights and Code of Conduct for Online Teaching

A misuse of online modes of communication is unacceptable. TAs and Faculty will seek consent before the recording of live online lectures or tutorials. Please ensure if you do not wish to be recorded during a session to inform the faculty member. Please also ensure that you prioritize formal means of communication (email, LMS) over informal means to communicate with course staff.