Carnegie Mellon University

94-879 Fundamentals of Operationalizing AI

Lecture Days, Times, Location: MW – 9:30AM-10:50AM; HBH 1002
Recitation Days, Times, Location: F – 2:00PM-3:20PM; HBH1005

Semester/Year: Fall 2024

Units: 6, **Section(s):** A1

Instructor information

Name	Prof Anand S Rao
Contact Info	anandr2@andrew.cmu.edu
Office location	Hamburg Hall 2105D
Office hours	Thursday – 2PM-3PM (Also available through Zoom)

Course Description

Artificial Intelligence (AI) is rapidly transforming industries by driving innovation, improving efficiency, and enhancing decision-making processes. Yet, despite its potential, many AI projects face significant hurdles in deployment. According to a recent survey, only 22% of data scientists report that their "revolutionary" AI initiatives—those designed to enable new processes or capabilities—usually reach deployment. Alarmingly, 43% say that 80% or more of their AI projects fail to make it into production. Even when considering all types of machine learning projects, including those focused on refreshing existing models, only 32% of models typically deploy. These statistics underscore the critical challenges of scaling AI systems effectively.

This course is designed to tackle these challenges head-on, providing graduate-level students with a comprehensive understanding of the AI lifecycle. Students will learn how to navigate the complex process of identifying which business tasks should be automated through AI and which decisions should be augmented using AI. The course introduces practical frameworks essential for making these strategic decisions and successfully implementing AI solutions.

Throughout the course, students will engage deeply with each stage of the AI lifecycle. They will learn to identify and prioritize high-impact AI use cases, conduct thorough cost-benefit analyses, and design strategic roadmaps aimed at maximizing return on investment (ROI). The curriculum blends theoretical knowledge with hands-on experience using industry-standard tools such as Jupyter Lab, Docker, Kubernetes, Kubeflow, Kafka, and Evidently. These tools are critical for overcoming the common pitfalls associated with AI deployment and preparing students to scale AI systems in real-world environments.

The course's practical orientation is further enhanced through case studies that serve as a foundation for class discussions. These case studies provide students with the opportunity to analyze real-world AI applications, assess the challenges involved, and understand the decision-making processes behind successful implementations.

Additionally, two guest lectures from seasoned industry practitioners will offer firsthand insights into the practical challenges of AI deployment across various sectors.

A strong emphasis is placed on governance and trust, equipping students with the knowledge to develop ethical, transparent, and effective AI systems. Students will learn how to integrate AI into organizational processes, assess talent and skill gaps, and create strategies to build the necessary capabilities for sustained AI-driven innovation.

This course is essential for students aspiring to careers as AI engineers, AI analysts, or AI governance experts. It is equally invaluable for business and technology students who wish to understand how to manage the development and deployment of AI systems. By the end of the course, students will possess a well-rounded, practical understanding of AI system management, enabling them to lead AI-driven projects and drive innovation across industries.

Learning Objectives

1. **AI Lifecycle Mastery:** Develop a thorough understanding of the AI system lifecycle, including the identification of business tasks for automation or augmentation, and effectively manage data preparation, model development, deployment, and maintenance to ensure alignment with organizational goals and industry best practices.
2. **Strategic Value Realization:** Identify and assess business needs, conduct detailed cost-benefit analyses, and design strategic roadmaps that prioritize high-impact AI initiatives. Develop AI models that deliver maximum ROI and are closely aligned with business objectives.
3. **Effective AI Operationalization:** Design and implement strategies to successfully deploy, monitor, and maintain AI models at scale, with a focus on overcoming common deployment challenges, ensuring system reliability, building trust, and fostering continuous improvement.
4. **Ethical Governance and Risk Management:** Establish and enforce comprehensive governance frameworks that promote ethical, transparent, and effective AI operations. Address potential risks proactively while fostering stakeholder trust and confidence in AI systems.
5. **Organizational Integration and Skill Building:** Evaluate organizational talent and process needs, and create actionable strategic plans to address skill gaps. Integrate AI seamlessly into business processes and build the necessary capabilities to sustain AI-driven innovation within the organization.
6. **Practical Tool Proficiency:** Gain hands-on experience with key AI lifecycle tools—Jupyter Lab, Docker, Kubernetes, Kubeflow, Kafka, and Evidently—preparing you to effectively scale and deploy AI systems in real-world environments.

Learning Resources

The following textbooks will be used as reference for the topics covered. Each topic will also have selected reading materials. The recommended book for purchase is:

1. Designing Machine Learning Systems by Chip Huyen, O'Reilly Media, May 2022.

In addition, the following books also cover similar topics to this course.

2. Operationalizing AI: How to accelerate and scale across people, process, and platforms by John J. Thomas, William Roberts, and Paco Nathan, O'Reilly Media, March 2021.
3. Operating AI by Ulrika Jagare, Wiley, May 2022.

Assessments

The final course grade will be calculated using the following categories:

Assessment	Percentage of Final Grade
Generative AI Assignment	3%
Class Participation	7%
Individual Assignment	30%
Three Class/Online Quizzes	30%
Team Project Presentation	30%
Total	100%

- **Generative AI Assignment:** More details on this will be provided during the first week of the course.
- **Class Participation:** Class participation would be based on (a) Coming prepared to the class having read the pre-reads; (b) Meaningful contributions to the case discussions and insightful questions during the lectures.
- **Individual Assignment:** Individual assignment will be based on answering the discussion questions in the industry-based synthetic scenarios or a programming assignment. See note on the use of generative AI tools in the Generative AI Guidance section.
- **Quizzes**: Three classroom/online quizzes will be administered during Week 2 to Week 6 (inclusive of both these weeks) of the course. Students are NOT allowed to use any AI tools or textbooks for the quizzes.
- **Team Project Presentation:** The final project will be a team presentation based on an industry-based synthetic scenario that will be distributed at the start of the course. The students will work as a team during the course and will make the final presentation to a panel of judges. There will be no final exam and the presentation will be conducted during the week of the exams.

Students will be assigned the following final letter grades, based on calculations coming from the course assessment section.

Grade	Percentage Interval
A+	98.0-100%
A	92.0-97.9%
A-	90.0-91.9%
B+	88.0-89.9%
B	82.0-87.9%
B-	80.0-81.9%
C+	78.0-79.9%
C	72.0-77.9%
C-	70.0-71.9%

D	50.0-69.9%
F	00.0-49.9%

Grading Policies

- **Late-work policy**: To encourage timely submissions and ensure fair and prompt grading for all students, assignments should be submitted by 11:59 PM on the due date. For those facing unforeseen circumstances, assignments may be submitted up to 24 hours late for up to 90% of the original grade, with incremental reductions thereafter. No assignments will be marked after 10 days.
- **Make-up work policy**: To maintain the integrity of the grading process while offering flexibility, there will be no make-up assignments or quizzes.
- **Re-grade policy**: To uphold the integrity of the assessment process, regrading will not be available. However, students are welcome to discuss the rationale for their grades during office hours to gain a better understanding of the assessment.
- **Attendance and/or participation policy**: To emphasize the value of class participation and active engagement in the learning process, attendance is mandatory and will be tracked via a sign-in sheet. Students have the flexibility to miss one class without affecting their class participation grade, as outlined in the Class Participation guidelines.

Course Policies

- **Academic Integrity & Collaboration**: Students are expected to strictly follow Carnegie Mellon University rules of academic integrity in this course. This means that unless otherwise specified, Individual assignments are to be the work of the individual student using only permitted material and without any cooperation of other students or third parties. It also means that usage of work by others is only permitted in the form of quotations and any such quotation must be distinctively marked to enable identification of the student's own work and own ideas. All external sources used must be properly cited, including author name(s), publication title, year of publication, and a complete reference needed for retrieval. The same work may not be submitted for credit in multiple courses. Violations will be penalized to the full extent mandated by the CMU policies. There will be no exceptions.
- **Use of Generative AI Tools**: We encourage students to explore the use of generative artificial intelligence (AI) tools, such as ChatGPT, for all individual assignments. Any such use must be appropriately acknowledged and cited, following the guidelines established by the APA Style Guide, including the specific version of the tool used. Submitted work should include the exact prompt used to generate the content as well as the AI's full response in an Appendix. Because AI generated content is *not* necessarily accurate or appropriate, it is each student's responsibility to assess the validity and applicability of any generative AI output that is submitted. You may not earn full credit if inaccurate, invalid, or inappropriate information is found in your work. Deviations from these guidelines will be considered violations of CMU's academic integrity policy.
- **Disabilities**: If you have a disability and have an accommodations letter from the Disability Resources office, I encourage you to discuss your accommodations and needs with me as early in the semester as possible. I will work with you to ensure that accommodations are provided as appropriate. If you suspect that you may have a disability and would benefit from accommodations but are not yet registered with the Office of Disability Resources, I encourage you to contact them at access@andrew.cmu.edu.
- **Student wellness**: As a student, you may experience a range of challenges that can interfere with learning, such as strained relationships, increased anxiety, substance use, feeling down, difficulty concentrating and/or lack of motivation. These mental health concerns or stressful events may diminish your academic performance and/or reduce your ability to participate in daily activities. CMU services are available, and treatment does work. You can learn more about confidential mental health services available on campus at

http://www.cmu.edu/counseling . Support is always available (24/7) from Counseling and Psychological Services: 412-268-2922.

- **Diversity:** It is my intent that students from all diverse backgrounds and perspectives be well served by this course, that students' learning needs be addressed both in and out of class, and that the diversity that students bring to this class be viewed as a resource, strength, and benefit. It is my intent to present materials and activities that are respectful of diversity: gender, sexuality, disability, age, socioeconomic status, ethnicity, race, and culture. Your suggestions are encouraged and appreciated. Please let me know ways to improve the effectiveness of the course for you personally or for other students or student groups.

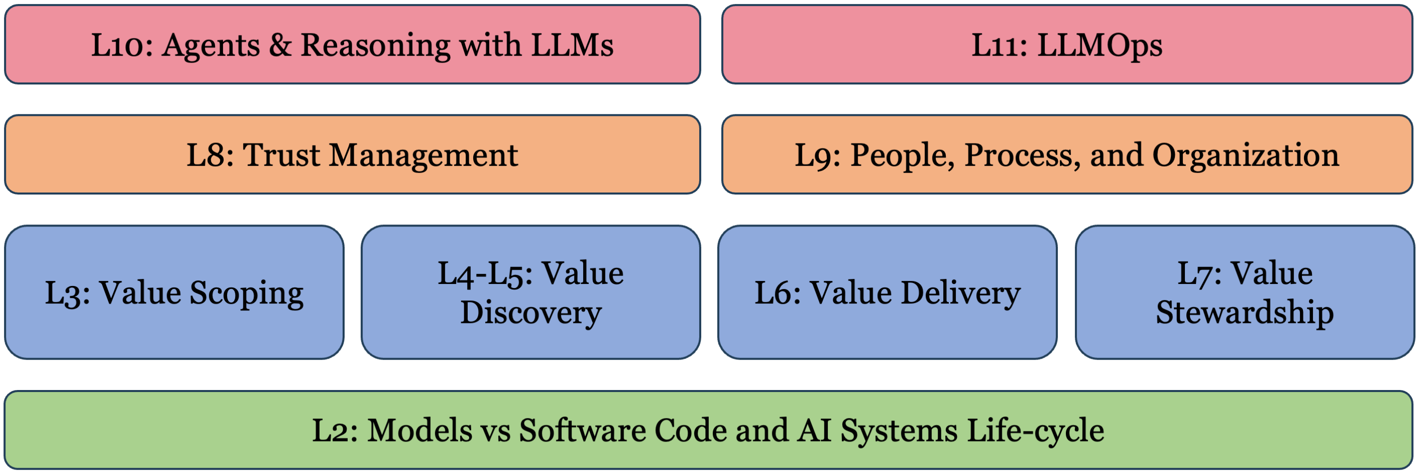
Course Schedule

Date	Theme/Topic	Learning Outcomes Addressed	Assignments Due
M: Aug 26	L1: Introduction & Overview	Clear understanding of the scope of the course, learning objectives, topics covered, student evaluation, and relevance of course to broader practice. *Case Study: Scaling AI*	
W: Aug 28	L2: Models vs Software code and AI Systems Life Cycle	Appreciate the differences between machine learning models and software code, implications of these differences and the consequences for AI systems management. Clearly explain the key components of the AI lifecycle, its maturity levels, and how they apply to various industries and functional areas. *Case Study: Data Scientists are from Mars & Software Developers are from Venus*	(Ungraded) Case Study Assignment -1
F: Aug 30	Recitation	Introduction to Shell, Anaconda, and Hugging Face environments	(Ungraded) Quiz-L1&L2
M: Sep 2	No Class	Labor Day Holiday	
W: Sep 4	L3: Value Scoping	Define and measure Key Performance Indicators (KPIs) for AI projects, calculate Return on Investment (ROI) *Case Study: ROI of AI: Navigating the road to value realization*	(Ungraded) Case Study Assignment - 2
F: Sep 6	Recitation	Introduction to Kafka Stream	(Ungraded) Quiz-L3&L4

		processing and Faust python package for stream processing	
M: Sep 9	L4: Value Discovery	Class quiz; Exploratory data analysis and feature stores	(Graded) Quiz-1
W: Sep 11	L5: Value Discovery	Learn best practices for model selection and evaluation. *Case Study: AI transformation journey*	(Ungraded) Case Study Assignment -3
F: Sep 13	Recitation	Introduction to Feast, an open-source feature engineering tool	(Ungraded) Quiz-L5&L6
M: Sep 16	L6: Value Delivery	Analyze methods for deploying models at scale. *Case Study: Engineering stability in the face of unprecedented growth*	(Ungraded) Case Study Assignment -4
W: Sep 18	L7: Value Stewardship	Evaluate challenges, need and techniques for monitoring data, models, and software. *Case Study: Harnessing change for innovation*	(Ungraded) Case Study Assignment -5
F: Sep 20	Recitation	Introduction to Dockers, Kubernetes and containerization of models.	(Ungraded) Quiz-L7&L8
M: Sep 23	Guest Lecture (TBD)	Identify and solve challenges of applying operationalization in the industry.	(Graded) Individual Assignment Due
W: Sep 25	L8: Trust Management	Create an end-to-end, top-down governance framework that ensures the consistent, efficient, and effective operation of AI systems.	(Graded) Quiz-2
F: Sep 27	Recitation	Introduction to Kubeflow and end-to-end AI system life-cycle.	(Ungraded) Quiz-L9
M: Sep 30	L9: People, Process, Organization	Assess the existing talent and skills within the organization and develop a plan to fill any identified gaps.	
W: Oct 2	Guest Lecture (TBD)	Identify and solve challenges of applying operationalization in the	

		industry.	
F: Oct 4	Recitation	Introduction to Evidently, an open-source model monitoring tool.	(Ungraded) Quiz-L10
M: Oct 7	L10: Generative AI	Assess the changes to be made to AI life cycle for generative AI. *Case Study: Strategic pivot to Generative AI*	(Graded) Quiz-3
W: Oct 9	L11: LLMOps & Reflections	Assess current trends and the future of operationalizing AI and recap key learnings from course	(Ungraded) Quiz-L11&L12
F: Oct 11	Recitation/Final	TBD	Final class presentations or Final revision

Course Architecture



L10: Agents & Reasoning with LLMs

L11: LLMOps

L8: Trust Management

L9: People, Process, and Organization

L3: Value Scoping

L4-L5: Value Discovery

L6: Value Delivery

L7: Value Stewardship

L2: Models vs Software Code and AI Systems Life-cycle

Course Outline

This mini-course is planned for fourteen sessions of 1 hour 20 minutes each.

- **Lecture 1: Introduction & Overview**
  - **Topics**
    - Introductions
    - Structure of the class
    - Expectations
    - What is "Operationalizing AI"?
    - Why study "Operationalizing AI"?
  - **Book**:
    - [Huyen] Chapter 1: Overview of Machine Learning Systems
- **Lecture 2: Models vs Software Code and AI System Life-cycle**
  - **Topics**
    - Case Study Discussion: Bridging Two Worlds: The Trials and Triumphs of Integraing AI into Traditonal Sofware at FinSolutions
    - Differences between machine learning models and software code
    - Consequences of mistaking models for software
    - Software and Data Science Lifecycles
    - AI System Lifecyle
    - Introduction to MLFlow
  - **Book**:
    - [Huyen] Chapter 1: Overview of Machine Learning Systems
    - [Huyen] Chapter 2: Introduction to Machine Learning Systems Design
  - **Required Reading:**
    - *Data Scientists are from Mars and Software Developers are from Venus (Part 1). Anand Rao Towards Data Science. August 29, 2020.*
    - *Consequences of mistaking models for software (Part 2). Anand Rao. Towards Data Science. September 6, 2020.*
    - *Model Evolution: From Standalone Models to Model Factory (Part 3)*,Towards Data Science, September 13, 2020.
    - *Model Lifecycle: From ideas to value.* Anand Rao, Towards Data Science, September 26, 2020.
  - **Optional Reading:**
    - *Why are machine learning projects so hard to manage? Lukas Biewald, Medium, January 28, 2019*
    - *Create a common-sense baseline first. Rama Ramakrishnan, Medium. January 12, 2018.*
    - *A guide to different bias mitigation techniques in machine learning. Sourabh Mehta, Analytics India. April 2, 2022.*
    - *MLOps: Machine Learning Life Cycle*, ML4Devs
    - *MLOps: Definitions, tools and challenges.* Symeonidis G., et.al.,, arXiv:2201.00162v1, January 2022.
- **Lecture 3: Value Scoping**
  - **Topics**
    - Case Study Discussion: ROI of AI: Navigating the Road to Value Realization at MidWest Financial
    - Developing a business case
    - ROI for AI
  - **Book**:
    - [Jagare] Chapter 7: Achieving Business Value from AI
  - **Required Reading:**

        - *Solving AI's ROI problem is not that easy.* Anand Rao, Tech Effect, July 20, 2021.
        - *How a Portfolio approach to AI helps your ROI*. Anand Rao, Tech Effect, September 9, 2021.
- **Lecture 4: Value Discovery**
    - **Topics**
        - Model discovery steps
        - Data engineering
        - Feature engineering and feature stores
    - **Book**:
        - [Huyen] Chapter 3: Data Engineering Fundamentals
        - [Huyen] Chapter 4: Training Data
        - [Huyen] Chapter 5: Feature Engineering
    - **Optional Readings:**
        - [Jagare] Chapter 2: Data Engineering Focused on AI
- **Lecture 5: Value Discovery**
    - **Topics**
        - Model development
        - Model evaluation
    - **Book**:
        - [Huyen] Chapter 6: Model Development and Offline Evaluation
    - **Optional Readings:**
        - [
- **Lecture 6: Value Delivery**
    - **Topics**
        - Myths of model deployment
        - Types of model deployment
        - Batch and Online architectures for inference
        - Deployment strategies
        - Model performance tuning
        - Docker
    - **Book**:
        - [Huyen] Chapter 7: Model Deployment and Prediction Service
- **Lecture 7: Value Stewardship**
    - **Topics**
        - Data drift and model drift
        - Continual learning
        - Infrastructure and tools for MLOps
    - **Book**:
        - [Huyen] Chapter 8: Data Distribution Shifts and Monitoring
        - [Huyen] Chapter 9: Continual Learning and Test in Production
        - [Huyen] Chapter 10: Infrastructure and Tooling for MLOps
    - **Optional Readings:**
        - *From concept drift to model degradation: An overview on performance-aware drift detectors. Bayram., et. Al., Knowledge- Based Systems 245, 2022.*
- **Guest Lecture: Operationalizing AI** – TBD

- **Lecture 8: Trust Management**
    - **Topics**
        - NIST RAI Framework
        - Bias and Fairness
        - Explainability and Interpretability
        - AI Governance
    - **Book**:
        - [Jagare] Chapter 6: AI is All About Trust
    - **Additional Readings:**
        - *Five Views of AI Risk: Understanding the darker side of AI (Towards Responsible AI — Part 1)*, Anand Rao. Towards Data Science, November 28, 2020.
        - *Ten principles of Responsible AI for corporates* (Towards Responsible AI — Part 2), x Anand Rao. Towards Data Science, December 17, 2020.
        - Top-down and end-to-end governance for the responsible use of AI (Towards Responsible AI — Part 3). Anand Rao. Towards Data Science, January 19, 2021.
        - Six stage gates to a successful AI Governance (Towards Responsible AI – Part 4). Anand Rao. Towards Data Science, February 21, 2021.
- **Lecture 9: People, Process, and Organization**
    - **Topics**
        - Emerging skills and capabilities
        - Operating models for AI
        - Center of Excellence
        - Human-centered AI
    - **Readings**:
        - [Thomas] Chapter 5: AI Center of Excellence
        - [Thomas] Chapter 2: Personas and Effective Communication Among Them
        - [Thomas] Chapter 3: Design Thinking
        - [Huyen] Chapter 11: The Human side of Machine Learning
- **Guest Lecture: Lecture on enterprise deployment and monitoring**
    - **Topics**
        - TBD
- **Lecture 10: Generative AI**
    - **Topics**
        - Generative AI
        - Transformer Models – Large Language Models
        - Prompt Engineering
        - Model Adaptation
        - Model Evaluation
    - **Readings**:
        - An overview of Large Language Models (LLMs) by Mostafa Ibrahim, Weights and Biases, June 22, 2023.
        - Essential guide to foundation models and large language models by Babar M. Bhatti, Medium, February 5, 2023.
        - Exploring opportunities in the Generative AI value chain by Tobias Härlin, Gardar Björnsson Rova, Alex Singla, Oleg Sokolov, and Alex Sukharevsky, McKinsey Digital, April 2023.
    - **Additional Readings & Resources**:
        - Prompt Engineering Guide

        - Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., . . . Wen, J. (2023). A Survey of Large Language Models. ArXiv. /abs/2303.18223
        - Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., . . . Gui, T. (2023). The Rise and Potential of Large Language Model Based Agents: A Survey. *ArXiv*. /abs/2309.07864.

- **Lecture 11: LLMOps**
    - **Topics**
        - Challenges of LLMOps
        - Best practices of LLMOps
        - Tools and frameworks for LLMOps
        - Cost and business case for pre-training, prompt engineering, and fine-tuning
        - Recap and summary
    - **Readings**:
        - Understanding LLMOps: Large Language Model Operations by Leonie, April 2023, Weights & Biases.
        - Emerging architectures for LLM Applications by Matt Bornstein and Rajko Radovanovic, Andreesen Horowitz, June 20, 2023.
    - **Additional Readings & Resources**:
        - What is LLMOps? Large Language Models' Ops, Architecture & Recommended tools by Arun, April 2023, Accubits.
        - A Developer's Guide To LLMOps: MLOps for Operationalizing LLMs by Hakan Tekgul and Aparna Dhinakaran, Arize.