# ECE595: AI Infrastructure – Spring 2H

Ioannis Papapanagiotou
ipapapa@unm.edu
Office hours will be announced through Canvas and held via Zoom

Electrical & Computer Engineering
MSC01 1100, 1 University of New Mexico
Albuquerque, NM 87131-0001
Department Phone: (505) 277-2436

## Course Description

This course provides a comprehensive overview of the infrastructure and technologies required to build, deploy, and manage AI systems, with a focus on Large Language Models (LLMs). Students will gain a deep understanding of the AI workflow, from data acquisition and model training to deployment and monitoring. The course will cover various aspects of AI infrastructure, including hardware, software, cloud platforms, and operational best practices.

## Course Goals

- Understand the fundamental components and workflow of AI systems.
- Learn how to build and deploy machine learning models in production environments.
- Explore the capabilities and challenges of LLMs and their applications.
- Gain practical experience with AI infrastructure tools and technologies.
- Develop skills in managing and operating AI systems in a secure and responsible manner.

## Student Learning Outcomes

The following are the Student Learning Outcomes for the course. Each module will have specific learning objectives listed on the Overview Page. The activities in that module (i.e.: discussions, assignments, and assessments) are developed so that you can demonstrate you have met these objectives:

- C1: Explain the AI Infrastructure components, understand the AI workflows and be able to architect an AI systems
- C2: Demonstrate how to run AI systems in production with common frameworks based on MLOps and AIOps
- C3: Build applications that leverage Generative AI, Large Language Models (LLMs);
- C4: Decide when to use small sized models like sLLMs as well as Multi-modal capabilities like Large Multimodal Models (LLMs)

## Course Map

| Modules | Lessons | Videos | Tentative Homework Schedule | Work and Evaluation | Learning Objectives |
|---|---|---|---|---|---|

Last update: 10/10/2024

| Module 1: Introduction to AI Infrastructure (week 1) | | | | | |
|---|---|---|---|---|---|
| 1.1 | Intro to AI Infrastructure | Intro to AI Infrastructure | N/A | Module 1 Quiz 1 (graded) | Define AI infrastructure components and their roles in AI systems.<br><br>Explain the workflow of an AI system, including data ingestion, preprocessing, model training, and deployment.<br><br>Architect an AI system based on given requirements and constraints.<br><br>Identify common challenges and best practices in AI infrastructure design. |
| 1.2 | | AI Workflow | | | |
| 1.3 | | AI Components | | | |
| 1.4 | | AI Compute | | Module 1 Quiz 2 (graded) | |
| 1.5 | | AI Application Frameworks | | | |
| 1.6 | | Cloud vs On Prem AI Infrastructure | | | |
| **Module 2: ML Infrastructure (weeks 2)** | | | | | |
| 2.1 | ML Infrastructure | ML Pipeline Introduction | Homework #1 | Module 2 Quiz 1 (graded) | Explain the components of an ML infrastructure, including data pipelines, model training environments, and deployment platforms.<br><br>Implement ML pipelines using tools like Kubeflow or Airflow.<br><br>Implement MLOps practices for |
| 2.2 | | ML Model Building | | | |
| 2.3 | | Data Challenges | | | |

| | | | | | |
|---|---|---|---|---|---|
| 2.4 | | Introduction to MLOps | | Module 2 Quiz 2 (graded) | version control, monitoring, and reproducibility. |
| 2.5 | | ML Feature Stores | | | |
| 2.6 | | ML Model Stores | | | |
| **Module 3: Generative AI** | | | | | |
| 3.1 | Generative AI | Introduction to LLMs | Homework #2 | Reading Assignment: Transformer architecute | Explain the concept of Retrieval Augmented Generation (RAG) and its benefits in generative AI. |
| 3.2 | | Transformer Architecture | | Module 3 Quiz 1 (graded) | Describe the capabilities and limitations of Large Language Models (LLMs) like GPT and Gemini. |
| 3.3 | | Applications of the Transformer Architecture | | | Combine RAG, LLMs, and embedding models to create powerful generative AI applications. |
| 3.4 | | LLM Parameters | | | |
| 3.5 | | Retrieval Augmentation | | Module 3 Quiz 2 (graded) | |
| 3.6 | | Small LLMs | | | |
| 3.7 | | Embedding | | | |

| | | | | | |
|---|---|---|---|---|---|
| | | Models | | | |
| 3.8 | | Large Multimodal Models | | | |
| **Module 4: LLM Infrastructure** | | | | | |
| 4.1 | LLM Infrastructure | Data Layer | Homework #3 | Reading Assignment: Langchain | Describe the data requirements for training and fine-tuning LLMs, including data quality, quantity, and diversity.<br><br>Understand the different types of LLM architectures and their trade-offs.<br><br>Describe when to use LLMs vs sLLMs and how to apply Multimodal capabilities |
| 4.2 | | Model Layer | | Module 4 Quiz 1 (graded) | |
| 4.3 | | Deployment Layer | | | |
| 4.4 | | Interface Layer | | | |
| 4.5 | | Key Takeways | | | |
| 4.6 | | Model Gardens: AWS Bedrock vs Google Vertex AI | | Module 4 Quiz 2 (graded) | |
| 4.7 | | AWS Bedrock or AWS Sagemaker | | | |
| | | Homework #4 | | | |
| **Module 5: LLM Operations** | | | | | |
| 5.1 | LLM Operations | LLM Security | Homework #4 | Module 5 Quiz (graded) | Explain the key concepts and practices of LLMOps, including |

| | | | | | |
|---|---|---|---|---|---|
| 5.2 | | LLMOps | | | version control, monitoring, and reproducibility.<br><br>Discuss the security risks associated with LLMs and how to mitigate them.<br><br>Evaluate the ethical implications of LLMs and develop responsible AI practices. |
| 5.3 | | LLM in Production | | | |
| 5.4 | | LLM Hallucinations | | | |
| End of Course | | | | | |

## Course Modality

*Online Asynchronous*

## Prerequisites and Co-requisites

Basic understanding of artificial intelligence and machine learning concepts, familiarity with cloud computing and data management principles.

Required course (or equivalent):

- ECE 530 Cloud Computing

Optionally, the students could also take:

- ECE 517 Machine Learning

## Technical Skills

To participate and succeed in this class, you will need to be able to perform the following basic technical tasks:

- Use Canvas (help documentation located in “Help”>”UNM Canvas Help Site” link on left course menu, and also at Online Student Documentation ).
- Use email – including attaching files, opening files, downloading attachments
- Copy and paste within applications including Microsoft Office
- Open a hyperlink (click on a hyperlink to get to a website or online resource)
- Use Microsoft Office applications
    - Create, download, update, save and upload MS Word documents
    - Create, download, update, save and upload MS PowerPoint presentations
    - Create, download, update, save and upload MS Excel spreadsheets
    - Download, annotate, save and upload PDF files

- Use the in-course web conferencing tool (Zoom)
- Have basic experience and understanding of Cloud technologies like Virtual Machines
- Download and install an application or plug in through Unix command line
- Have knowledge and be able to write some basic Python code

### Technical Requirements

## Computer

- A high-speed Internet connection is highly recommended.
- Supported browsers include: Detailed Supported Browsers and Operating Systems.
- Any computer capable of running a recently updated web browser should be sufficient to access your online course. However, bear in mind that processor speed, amount of RAM, and Internet connection speed can **greatly** affect performance. Many locations offer free high-speed Internet access including UNM's Computer Pods.
- Microsoft Office products are available free for all UNM students (more information on the UNM IT Software Distribution and Downloads page).
- Access to a system that has a command line such as Mac OS or Linux.

**For UNM Canvas Technical Support: (505) 277-0857 (24/7) or visit the Canvas Info Site**

**Canvas outages: Unexpected Canvas system outages are rare but, if they occur, I will advise everyone on how to proceed.**

## Web Conferencing

Web conferencing will be used in this course during the office hours.

For the online sessions, you will need:

- Headphones and a computer mic OR a headset with microphone. Headsets are widely available at stores that sell electronics, at the UNM Bookstore, or online.
- A high-speed internet connection is highly recommended for these sessions. Please test your wireless Internet connection for audio and/or video quality prior to web conferencing.
- For UNM Web Conference Technical Help: (505) 277-0857

## Tracking Course Activity

Canvas automatically records all students' activities including: your first and last access to the course, the pages you have accessed, the number of discussion messages you have read and sent, web conferencing, discussion text, and posted discussion topics. This data can be accessed by the instructor to evaluate class participation and to identify students having difficulties in the class.

### Textbook and Supplemental Materials

The AI Infrastructure subject is novel and dynamic, so there is no recommended textbook. If you are interested in reading material, you can also look for subjects like MLOps or AIOps. The course instructor will provide reading material and online resources at the beginning of each section.

## Required Textbooks:

No required textbooks

## Recommended and/or Optional Textbooks, Journals and Articles:

- **Textbook:** AI Engineering: Building Applications with Foundation Models – Chip Huyen
- **Short Courses:** https://www.deeplearning.ai/ Most of them are introduced by Andrew Ng.
- **Podcast**: If you're on the move a lot or have a long commute, podcasts can be a great way to use that time to learn something new. There are many machine learning and MLOps podcasts out there. Most conduct interviews with MLOps professionals, which is a great way to peek behind the curtain of industry leaders' machine learning teams.
    - ML Platform Podcast
    - The MLOps Community podcast
    - TWIML
    - Practical AI
    - The Machine Learning Podcast

### Coursework and Participation

## Instructor Response Time

Course messages are checked within 48h Monday through Friday, and all messages will be responded to within 72 hours (about 3 days). If you do not hear from the instructor within 72 hours, please send your message again.

## Assignments

- For remote exams, we will use UNM Canvas
    - All written work needs to be submitted online. If you have difficulty using a tool to complete work, please reach out to UNM's Canvas Support immediately and notify your instructor as well.
- There will be no makeup exams or homework.
    - Late submissions will lose 10% of the corresponding grade of the submission per 24h.
    - If a student anticipates difficulty in meeting the deadline, please inform the instructor at least 72h before
- All written work needs to be submitted online. If you have difficulty using UNM Learn, please create a support ticket (using the course menu) and notify your instructor as well.

Last update: 10/10/2024

## Procedures for Completing Coursework

- If you start a Quiz or a test you must complete within the specified amount of time. If the Quiz is not completed, a new one will not be provided. Please make sure that you have enabled pop ups on your browser for Canvas. Test that your browser is functional before the quizzes
- All assignments must be delivered on time.
    - Late assignments will include a -10% penalty per day of delay. Late days begin to count at the first second post the deadline of each assignment.
    - If you cannot meet a deadline please work with the instructor at a minimum 48h before the begin date on the assignment.
- All work must be original. No LLM or AI tool should be used for the reading assignments.
- *All written work needs to be submitted online. If you have a difficulty using a tool to complete work, please reach out to UNM's Canvas Support immediately and notify your instructor as well."*

The course has reading assignments, homework and quizzes. Please refer to Canvas for the description and deadlines.

## Credit Hours, Expected Time Commitment, and Expectations for Participation

This is a three credit-hour course delivered in an entirely online modality over 8 weeks during the Spring 2025 semester. Please plan for a *minimum* of 18 hours per week to learn course materials and complete assignments.

**Additional Expectations:**

- students are expected to learn how to navigate in Canvas
- students are expected to communicate with one another in team assignments
- students are expected to keep abreast of course announcements
- students are expected to use the Canvas course inbox or UNM email as opposed to a personal email address
- students are expected to keep instructor informed of class related problems, or problems that may prevent the student from full participation
- students are expected to address technical problems immediately
- students are expected to observe course netiquette at all times

## Netiquette

Students are expected to follow the guidelines of netiquette when communicating and interacting in our course. Netiquette refers to a set of guidelines in online communication that help ensure positive interactions. In this case specifically, these guidelines seek to keep our online class a positive learning environment for everyone.

## Grading Procedures

- Hands on Lab/Homework (total 4 – 10% each): 40%
- Quizzes: 40%
- Reading Assignments: 20%

Feedback on the assignments will be returned within 10 business days.

## Grading Scale

### Undergraduate grading scale

| Letter Grade | Percentage | Letter Grade | Percentage |
|---|---|---|---|
| A+ | [97-100+ | C | [74-77) |
| A | [94-97) | C- | [70-74) |
| A- | [90-94) | D+ | [67-70) |
| B+ | [87-90) | D | [64-67) |
| B | [84-87) | D- | [60-64) |
| B- | [80-84) | F | 59 or Below |
| C+ | [77-80) | | |

### Graduate grading scale

| Letter Grade | Percentage | Letter Grade | Percentage |
|---|---|---|---|
| A+ | [97-100+ | B | [84-87) |
| A | [94-97) | B- | [80-84) |
| A- | [90-94) | C | [74-77) |
| B+ | [87-90) | F | [0-74) |
| C+ | [77-80) | | |

## UNM Policies

## Title IX

The University of New Mexico and its faculty are committed to supporting our students and providing an environment that is free of bias, discrimination, and harassment. The University's programs and activities, including the classroom, should always provide a space of mutual respect, kindness, and support without fear of harassment, violence, or discrimination. Discrimination on the basis of sex includes discrimination on the basis of assigned sex at birth, sex characteristics, pregnancy and pregnancy related conditions, sexual orientation and gender identity. If you have encountered any form of discrimination on the basis of sex, including sexual harassment, sexual assault, stalking, domestic or dating violence, we encourage you to report this to the University. You can access the confidential resources available on campus at the LoboRESPECT Advocacy Center, the Women's Resource Center, and the LGBTQ Resource Center. If you speak with an instructor (including a TA or a GA) regarding an incident connected to discrimination on the basis of sex, they must notify UNM's Title IX Coordinator that you shared an experience relating to Title IX, even if you ask the instructor not to disclose it. The Title IX Coordinator is available to assist you in understanding your options and in connecting you with all possible resources on and off campus. For more information on the campus policy regarding sexual misconduct and reporting, please see UNM Administrative Policy 2740 and CEEO's website.

If you are pregnant or experiencing a pregnancy-related condition, you may contact UNM's Office of Compliance, Ethics, and Equal Opportunity at ceeo@unm.edu. The CEEO staff will

provide you with access to available resources and supportive measures and assist you in understanding your rights. UNM's lactation stations are marked on the UNM campus map.

## Copyright Issues

All materials in this course fall under copyright laws and should not be downloaded, distributed, or used by students for any purpose outside this course.

The UNM Copyright Guide has additional helpful information on this topic.

## Accessibility

Accommodations: UNM is committed to providing equitable access to learning opportunities for students with documented disabilities. As your instructor, it is my objective to facilitate an inclusive classroom setting, in which students have full access and opportunity to participate. To engage in a confidential conversation about the process for requesting reasonable accommodations for this class and/or program, please contact Accessibility Resource Center at arcsrvs@unm.edu or by phone 505-277-3506.

## Academic Misconduct

You should be familiar with UNM's Policy on Academic Dishonesty and the Student Code of Conduct which outline academic misconduct defined as plagiarism, cheating, fabrication, or facilitating any such act.

## Example Drop Policy

This course falls under all UNM policies for the last day to drop courses, etc. Please see the UNM Course Catalog for information on UNM services and policies. Please see the UNM academic calendar for course dates, the last day to drop courses without penalty, and for financial disenrollment dates.

## UNM Resources

**CTL tutoring services** (previously known as CAPS) are free and available to UNM students enrolled in undergraduate classes. Services include tutoring in STEM, writing, and languages, and personalized support with study skills and learning strategies. For asynchronous feedback on your writing, submit assignments to the Online Writing Lab (OWL).

Services are available in person and online. Visit the main office on the 3rd floor of Zimmerman Library or go to the Virtual Front Desk on the CTL website to get connected with a tutor or to schedule an appointment: ctl.unm.edu >undergraduates.

**UNM Libraries** provides students with a number of ways to access research and resources, such as reserving books and other media, requesting books and online materials through ILLiad, and access to research databases.

**Student Health and Counseling** (SHAC) provides quality health and counseling services to all UNM students to foster student success. Fees charged at SHAC are much lower than

community rates. SHAC is funded in part by student fees, and they are accredited by the Accreditation Association for Ambulatory Healthcare (AAAHC). You can contact SHAC at (505) 277-3136.

**LoboRESPECT Advocacy Center** (505) 277-2911 can offer help with contacting faculty and managing challenges that impact your UNM experience.

## FOR MILITARY-CONNECTED STUDENTS

There are resources on campus designed to help you succeed. You can approach any faculty or staff for help with any issues you may encounter. Many faculty and staff have completed the GREEN ZONE training to learn about the unique challenges facing military-connected students. If you feel that you need help beyond what faculty and/or staff can give you, please reach out to the Veterans Resource Center on campus at 505-277-3181, or by email at vrc@unm.edu.

## LAND ACKNOWLEDGEMENT

Founded in 1889, the University of New Mexico sits on the traditional homelands of the Pueblo of Sandia. The original peoples of New Mexico Pueblo, Navajo, and Apache since time immemorial, have deep connections to the land and have made significant contributions to the broader community statewide. We honor the land itself and those who remain stewards of this land throughout the generations and also acknowledge our committed relationship to Indigenous peoples. We gratefully recognize our history.
Resource: Division for Equity and Inclusion.

### Citizenship and/or Immigration Status

All students are welcome in this class regardless of citizenship, residency, or immigration status. Your professor will respect your privacy if you choose to disclose your status. As for all students in the class, family emergency-related absences are normally excused with reasonable notice to the professor, as noted in the attendance guidelines above. UNM as an institution has made a core commitment to the success of all our students, including members of our undocumented community. The Administration's welcome is found on our website.

### Responsible Learning AND ACADEMIC HONesty

Cheating and plagiarism (academic dishonesty) are often driven by lack of time, desperation, or lack of knowledge about how to identify a source. Communicate with me and ask for help, even at the last minute, rather than risking your academic career by committing academic dishonesty. Academic dishonesty involves presenting material as your own that has been generated on a website, in a publication, by an artificial intelligence algorithm (AI), by another person, or by otherwise breaking the rules of an assignment or exam. It is a Student Code of Conduct violation that can lead to a disciplinary procedure. When you use a resource (such as an AI, article, a friend's work, or a website) in work submitted for this class, document how you used it and distinguish between your original work and the material taken from the resource.

Support: Many students have found that time management workshops or work with peer tutors can help them meet their goals. These and other resources, including support on how to cite a source, are available through Student Learning Assistance at the Center for Teaching and Learning.

Last update: 10/10/2024

## Connecting to Campus and Finding Support

UNM has many resources and centers to help you thrive, including opportunities to get involved in campus life, research experiences, mental health resources, academic support such as tutoring, resource centers for people like you, free food at Lobo Food Pantry, jobs on campus and financial capability support. Your advisor, staff at the resource centers and Dean of Students, and I can help you find the right opportunities for you.

## Respectful Conduct Expectations

I am committed to building with you a positive classroom environment in which everyone can learn. I reserve the right to intervene and enforce standards of respectful behavior when classroom conduct is inconsistent with University expectations [and/or classroom community agreements]. Interventions and enforcement may include but are not limited to required meetings to discuss classroom expectations, written notification of expectations, and/or removal from a class meeting. Removal from a class meeting will result in an unexcused absence. [Insert number] or more unexcused absences may result in permanent removal and a drop from the course (see attendance policy). The University of New Mexico ensures freedom of academic inquiry, free expression and open debate, and a respectful campus through adherence to the following policies: D75: Classroom Conduct, Student Code of Conduct, University Policy 2240 – Respectful Campus, University Policy 2210 – Campus Violence.