Master Course Description for EE-517

Title: Introduction to Large Language Models

Credits: 3

Course Catalog Entry:

EE 517: Introduction to Large Language Models is an introduction to the principles and applications of large language models (LLMs). Topics include the architecture, model training and fine-tuning, evaluation methods, ethical considerations, and real-world applications.

Coordinator: Banghua Zhu, Assistant Professor of Electrical and Computer Engineering

Goals

To provide students with a foundational understanding of large language models, including their architecture, training methods, and applications. To equip students with practical skills for working with LLMs and to prepare them for advanced research or industry work in generative AI.

Learning Objectives

At the end of this course, students will be able to:

1. Explain the key components of large language models.
2. Describe how LLMs are trained, fine-tuned, and adapted to specific tasks.
3. Implement and experiment with pre-trained LLMs using modern frameworks.
4. Evaluate the performance and limitations of LLMs using established benchmarks.
5. Identify and analyze ethical concerns related to bias, fairness, and misinformation in LLMs.
6. Apply LLMs to real-world tasks such as document AI, agentic tasks, .

Textbook

No single textbook; course materials will be drawn from research papers, online resources, and instructor-provided notes.

Prerequisites by Topic

1. Fundamentals of machine learning and deep learning.
2. Linear algebra, probability, and optimization.
3. Proficiency in Python.

Prerequisites by Course

1. Machine Learning: EE 344 or EE 345 or equivalent.
2. Python: EE 241 or CSE 163

Topics

1. Introduction to language models and self-supervised learning.
2. The transformer architecture and self-attention mechanism.
3. Pre-training, fine-tuning, and prompt engineering.
4. Evaluation metrics and benchmarking for LLMs.
5. Scaling laws and emergent properties of large models.
6. Efficiency techniques (quantization, distillation, retrieval-augmented generation).
7. Applications: text generation, translation, summarization, question-answering, coding assistants.
8. Ethical considerations: bias, fairness, misinformation, and environmental impact.
9. LLM deployment and real-world considerations.
10. Future directions and open challenges in LLM research.

Course Structure

Class meets twice per week for 80-minute lectures. The grading scheme in any particular offering is the prerogative of the instructor. Homework assignments include conceptual, coding exercises and a final project.

Grading

- Homework (40%)
- Course project (50%)
- Participation (10%)

Computer Resources

Students will require access to GPUs for assignments and projects. Cloud-based or department-provided resources will be available.

Laboratory Resources

Not required.

Preparers: Banghua Zhu
Revision Date: 03/01/2025