Course number and name: **CS 07552: Graduate Large Language Models**

Credits and contact hours: 3 credits / 3 contact hours

Instructor's or course coordinator's name: Sahana Varadaraju

Instructional materials:

1. Natural Language Processing with Transformers, Lewis Tunstall, Leandro von Werra, Thomas Wolf, O'Reilly Media, 2022
2. Transformers for Natural Language Processing, Denis Rothman (Packt) 2022
3. Open-source articles, tutorials, and arXiv research papers

Specific course information

Catalog description: This hands-on course introduces the architecture, mechanics, and application of Large Language Models (LLMs) through the lens of transformer-based systems like GPT, BERT, and T5. Students will gain practical experience in training, fine-tuning, and prompting LLMs for use cases such as chatbots, summarization, and retrieval-augmented generation. The course balances theoretical foundations with real-world projects, tools, and deployment strategies. Class will explore traditional issues including hallucination, safety, and governance of LLMs. A final project requires students to build and present an LLM-powered system.

Prerequisites: Python programming proficiency. Familiarity with linear algebra and probability helpful. Prior exposure to machine learning or deep learning recommended.

Specific goals for the course

1. **LLM Architecture.** Students will be able to describe key architectural components of transformer-based LLMs.

2. **Natrual Language Processing**. Students will pretrain, fine-tune, and prompt LLMs for various NLP tasks.

3. **Performance Evaluation.** Students will evaluate LLM performance using standard benchmarks.

4. **LLM Development.** Students will develop a working LLM-powered chatbot using public APIs or open-source models.

5. **Social Issues.** Students will understand ethical, privacy, and safety implications of generative models

List of topics to be covered:

1. Introduction to LLMs and NLP evolution
2. Word embeddings and tokenization methods
3. Transformer architecture and attention
4. GPT/BERT models and pre-training strategies
5. Prompt engineering and sampling techniques
6. Instruction tuning and Reinforcement Learning with Human Feedback (RLHF)
7. Retrieval-Augmented Generation (RAG) pipelines and fine-tuning techniques
8. Model evaluation
9. LLM-powered agent systems and tools (LangChain)
10. Case studies and applications across domains
11. Ethical considerations: hallucinations, bias, misuse
12. Advanced topics: Multimodal LLMs, foundation models