**Syllabus:**
**Generative AI - Foundations, Techniques, Challenges, and Opportunities**

## Course Description:

Recent developments in neural network architectures, algorithms, and computing hardware have led to a revolutionary development usually referred to as Generative AI nowadays. Large Language Models (LLMs) are now able to generate seemingly human-like text in response to tasks like summarization, question answering, etc with high level of accuracy. Leveraging similar strategies, comparable advances have been made with images as well as audio. With today's (and anticipated future) capabilities, Generative AI is poised to be a tool used comprehensively in a wide variety of ways, and therefore to have a profound set of effects on our lives and society as a whole.

This course is a broad introduction to these new technologies. It is split conceptually into three parts. In the Introduction section we will cover the historical aspects, key ideas and learnings all the way to Transformer architectures and training aspects. In the Practical Aspects and Techniques section, we will learn how to deploy, use, and train LLMs. We will discuss core concepts like prompt tuning, quantization, and parameter efficient fine-tuning, and we will also explore use case patterns. Finally, we will discuss challenges & opportunities offered by Generative AI, where we will highlight critical issues like bias and inclusivity, fake information, and safety, as well as some IP issues.

Our focus will be on practical aspects of LLMs to enable students to be both effective and responsible users of generative AI technologies.

## Course Goals and Objectives

By the completion of this course, students will:

- Appreciate the history of the path towards Large Language Models (LLMs) and Generative AI approaches.
- Understand the foundations of LLMs, how they are trained, and how to deploy and use them, for and beyond text-focused problems.
- Be able to understand key use case patterns of Generative AI approaches and know how to think about incorporating them into applications.
- Become conversant in PyTorch and key neural net coding strategies.
- Know how to approach improving the results obtained from LLMs through prompt-tuning, instruction-based fine-tuning, and Reinforcement Learning with Human Feedback.
- Become aware of critical issues such as bias, inclusivity problems, hallucinations, and IP questions

**Course Structure**

- The course will consist of weekly async material and weekly live sessions.
- There will be 4 homeworks and one final assignment. The homeworks will each count 5-20% for the grade, depending on complexity. The final Assignment, for which you will have 4 weeks, will count 40%.
- In addition, we will offer optional/additional material that will be taught live (and recordings will be made available):
    - Covering necessary background material:
        - 3 lectures/study meetings - 'Review of Neural Nets', 'Intro to PyTorch', and 'PyTorch and Transformers' - will be taught live for each interested student who can attend (optional), and then be made available as async content. The material is then considered must-read async material
    - Helping interested students to get to the forefront of current research (optional):
        - 1 lecture 'Deep Dive into Transformers', covering effectively some material that is otherwise taught in MIDS 266
        - About 4 paper reading sessions.

  Attendance of these components is strictly optional, and the content will not be assumed for homeworks or other parts of the class.

Make sure you give yourself enough time to be successful! In particular, you will be in for a rough semester if you have other significant commitments at work or home, or take both this course and any of MIDS 210 (Capstone), MIDS 261, MIDS 266, or MIDS 271.

**Grading**

The grading will be based on the Homework (60% combined) and the Final Assignment (40%).

**Course Prerequisites & Technical Information**

- MIDS 207 (Applied Machine Learning): We assume knowledge of all aspects of MIDS 207, particularly of the Neural Network architectures and their training as taught in that course.
- Strong coding capabilities in Python are a prerequisite. All assignments and notebooks will be based on Python. All assignments require coding.
- PyTorch will be our core framework. Prior familiarity is a plus, but not required. A targeted overview of PyTorch will be provided in the early part of the course. (You can use TensorFlow if you choose to do so.)

**Costs Incurred**

- Some of the assignments may require the use of Google Pro and their better GPUs and increased memory. The cost of this service is $10/month. You may incur as much as $50 over the course of the semester.

**Weekly Live Presentation of Lectures (replacing async material):**

- All Sections: Friday 4:00 - 5:30 pm PST

**Weekly Live Section Sessions:**

- Section 1: Tuesday 4:00 - 5:30 pm PST (Mark Butler)
- Section 2: Tuesday 6:30 - 8:00 pm PST (Joachim Rahmfeld)
- Section 3: Wednesday 6:30 - 8:00 pm PST (Mark Butler)
- Section 4: Thursday 6:30 - 8:00 pm PST (Joachim Rahmfeld)

# Deliverables

| Number | Topic | Released |
|---|---|---|
| Assignment 1 | **PyTorch Basics, HuggingFace, and a Simple Application: Sentence Classification**<br><br>This assignment will have two weeks for completion and will be worth **15%** of their grade. | End of Week Week 2 |
| Assignment 2 | **GPT-2: Evaluation of pre-trained and fine-tuned models**<br><br>This assignment will have two weeks for completion and will be worth **15%** of their grade. | End of Week Week 4 |
| Assignment 3 | **Image Generation and evaluation**<br><br>This assignment will have one week for completion and will be worth **10%** of their grade. | End of Week Week 6 |
| Assignment 4 | **Prompt Engineering**<br><br>This assignment will have two weeks for completion and will be worth **20%** of their grade. | End of Week Week 7 |
| Final Assignment | **Building a Retrieval-Augmented Q&A System**<br><br>This assignment will have four weeks for completion and will be worth **40%** of their grade. | End of Week Week 9 |

# Course Schedule/Syllabus

| Week | Title | Covered | Readings |
|---|---|---|---|
| | **Part I: Introduction** | | |
| 1 | **How did we get here? And where are we, really?** | Course overview<br><br>History of AI<br><br>AI & Neural Nets<br><br>Neural Nets & Language<br><br>Language and Reasoning<br><br>LLMs as a Black Box...<br><br>Are we done? | • TBD |
| 2 | **At the heart of it all: Context, context, context... (in language)** | the importance of context a first look:<br><br>- word embeddings trained form context<br><br>- ANYTHING can be represented as a vector given context<br><br>- limitations of word embeddings -> RNNs<br><br>- limitations of RNNs -> Transformers | • TBD |
| 3 | **LLMs I: Usage Patterns, Pre-training & Fine-Tuning** | Pre-training<br><br>Fine-tuning<br><br>In-context learning | • TBD |
| 4 | **LLMs II: Reinforcement Learning & RLHF, and keeping LLMs on task** | Reinforcement Learning/RLHF<br><br>Instruction-based Fine Tuning<br><br>Alignment | • TBD |

| | | | |
|---|---|---|---|
| 5 | **Context & Transformers: usages beyond NLP** | Encoder & Decoder architectures (high-levelish)<br><br>Pre-training in NLP (BERT & GPT)<br><br>Transformers for Vision & Audio,<br><br>Mixed models (CLIP, CLAP...)<br><br>Diffuson Models | • TBD |
| | **Part II: Practical Aspects & Techniques** | | |
| 6 | **Model & Training Efficiencies: Quantization, QLoRA, LoRA, Adapters and all that** | Distillation<br><br>LoRa<br><br>Quantization methods and QLoRa<br><br>Soft prompts & Adapters | • TBD |
| 7 | **Prompt Engineering** | Building prompt intuition<br><br>Basic prompt construction<br><br>Advanced prompt construction<br><br>Complex prompt structures<br><br>Automated prompt construction | • TBD |
| 8 | **Advanced Topics (context length, MoE, Grouped Attention)** | Techniques for extending context length<br><br>Strategies for using longer context<br><br>Mixture of Experts models<br><br>Grouped attention | • TBD |
| 9 | **Using LLMs: Deployment Options & Operations** | Hosted ecosystem: OpenAI, Cohere, Anthropic…<br><br>HuggingFace<br><br>GCP, AWS, Azure deployments | • TBD |

| | | | |
|---|---|---|---|
| 10 | **Usage Patterns I: Retrieval-augmented Q&A and More** | Semantic Search<br><br>Retrieval Augmentation<br><br>LLama Index<br><br>Hallucination issues | ● TBD |
| 11 | **Usage Patterns II: Chaining & Agents** | Langchain<br><br>React<br><br>Toolformer | ● TBD |
| 12 | **Multi-modal Large Language Models: Text & Language, Images, and Sound** | Training of Image/Language Transformers<br><br>Text-to-image and image-to-text models<br><br>Sound to Text<br><br>Dual input models<br><br>Sound and Generative AI | ● TBD |
| | **Part III: Challenges & Opportunities** | | |
| 13 | **Bias & Inclusivity, Safety, and Evaluation Approaches** | Safety metrics and evaluations<br><br>Bias questions<br><br>Faithfulness to facts vs making stuff up<br><br>Deep Fakes and their potential impact | ● TBD |
| 14 | **Some societal and legal considerations: Deep Fakes, usage rights, and IP questions, and the HUGE potential of LLMs & AI** | The fine-print in hosted models<br><br>The issue of hosted models learning from your data<br><br>IP questions around LLM... and their impact<br><br>The future of LLMs - a discussion | ● TBD |