Carnegie Mellon University

95-820 Applications of NL(X) and LLM

Lecture Days, Times, Location: MW – 3:30PM- 4:50PM; HBH 2003
Recitation Days, Times, Location: F – 3:30PM- 4:50PM; HBH 2003

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 – 3PM-4PM (Also available through Zoom)

Course Description

The rapid adoption of Generative AI technologies is reshaping industries across the globe, from healthcare to financial services. According to Bain & Company, 87% of companies were already developing, piloting, or deploying generative AI by the start of 2024. Moreover, McKinsey's research estimates that generative AI could contribute an additional $2.6 trillion to $4.4 trillion annually across various sectors. These figures underscore the urgent need for professionals equipped with specialized knowledge in Natural Language Processing and Understanding (NL(X)) and Large Language Models (LLMs).

This course is designed to provide graduate-level students with a comprehensive understanding of NL(X) and LLMs, focusing on their applications, evaluation, and operationalization across diverse industries. Beginning with the fundamentals of NL(X), the course covers its history, evolution, and critical applications, offering students hands-on experience with essential tools such as text mining, sentiment analysis, and embeddings.

As students' progress, they will delve into advanced architectures, including RNNs, LSTMs, and Transformers, learning how these models drive key applications like machine translation and named entity recognition (NER). The course places significant emphasis on Large Language Models, such as GPT and BERT, guiding students through the intricacies of training, fine-tuning, and deploying these models. Advanced topics like Retrieval-Augmented Generation (RAG) and agentic architectures will also be explored, highlighting how LLM-based agents are transforming tasks that require complex reasoning, planning, and execution.

To bridge the gap between theory and practice, the course offers detailed instruction on LLMOps, covering best practices for transitioning models from development to production. This includes a strong focus on ethical considerations, operational risks, and the optimization of model performance in enterprise settings.

Students will also benefit from guest lectures by industry professionals, providing valuable insights into the practical challenges and opportunities of applying NL(X) and LLMs in real-world environments. These sessions are designed to help students connect their academic learning with industry needs, preparing them to lead in the fast-evolving field of AI.

Given the accelerating adoption of Generative AI, this course is essential for those aspiring to roles as NLP engineers, data scientists, AI analysts, or professionals looking to leverage LLMs to drive innovation. By the end of the course, students will possess the critical skills and knowledge required to develop, evaluate, and deploy advanced NL(X) and LLM solutions, positioning themselves at the forefront of AI-driven transformation.

Learning Objectives

Upon completion of this course, students will be able to:

1. **Foundations of Natural Language Processing (NLP):** Develop a deep understanding of the history, evolution, and taxonomy of NLP, with a focus on its expanding role in generative AI technologies. Gain practical experience with essential NLP tools and techniques, including text mining, sentiment analysis, and TF-IDF, to address industry-specific challenges in healthcare, financial services, and beyond.
2. **Deep Learning and Embeddings in NLP:** Master the fundamentals of deep learning as applied to NLP, including the development and training of neural networks and the use of word and document embeddings. Apply these techniques to real-world industry applications, addressing the growing demand for expertise in models such as Word2Vec, Doc2Vec, RNNs, LSTMs, and GRUs.
3. **Advanced NLP Architectures and Techniques:** Gain proficiency in advanced NLP models, including Sequence-to-Sequence models, attention mechanisms, and Transformer-based architectures. Explore their critical role in generative AI applications such as machine translation, chatbots, and text classification, and learn to implement and fine-tune large language models (LLMs) like GPT, BERT, and their variants for domain-specific tasks.
4. **Evaluation and Benchmarking of LLMs:** Understand the importance of and techniques for benchmarking large language models (LLMs) using standard metrics and datasets. Develop the ability to navigate the practical challenges of evaluating LLMs in real-world scenarios, ensuring their ethical deployment and optimal performance in industry applications.
5. **Enterprise Applications and Retrieval-Augmented Generation (RAG):** Explore the growing enterprise applications of LLMs, with a focus on Retrieval-Augmented Generation (RAG) and its transformative impact on real-world tasks such as recommendation systems. Learn to customize embeddings and model architectures to meet the specific needs of various industries, enhancing performance and scalability.
6. **Agentic Architectures and LLM-based Agents:** Understand the principles and applications of agentic architectures powered by LLMs, focusing on their use in complex reasoning, planning, and multi-modal tasks. Explore how LLM-based agents are revolutionizing fields such as human behavior simulation, urban planning, and game theory, driving innovation across diverse industries.
7. **LLMOps and Ethical Deployment:** Master the best practices for developing, deploying, and operating NLP and LLM applications in production environments. Address the ethical challenges and potential risks in NLP and gain hands-on experience with deploying both closed-source and open-source LLMs, ensuring cost-effective and scalable solutions that meet the growing industry demand for generative AI.

These learning objectives are designed to provide students with a comprehensive understanding of both the foundational and cutting-edge concepts in NLP and LLMs, as well as practical skills for implementing, evaluating, and deploying these models in real-world applications, including the emerging area of LLM-based agents.

This course requires a basic background in data science and/or Artificial Intelligence. Basic level of Python programming and deep learning fundamentals is required for completing the assignments.

Learning Resources

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

1. Real-World Natural Language Processing by Masato Hagiwara, Manning Publications, November 2021.
2. Quick Start Guide to Large Language Models: Strategies and Best Practices for Using ChatGPT and Other LLMs by Sinan Ozdemir, Addison-Wesley Professional, October 2023.

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

3. Natural Language Processing with Transformers by Tunstall, L., Leandro von Werra, Thomas Wolf, O'Reilly Online Learning, May 2022.
4. Developing Apps with GPT-4 and ChatGPT by Olivier Caelen, Marie-Alice Blete, O'Reilly Media, Inc. August 2023.
5. Natural Language Understanding with Python by Deborah A. Dahl, Packt Publishing, June 2023.
6. Deep Learning for Natural Language Processing By Karthiek Reddy Bokka, Shubhangi Hora, Tanuj Jain, Monicah Wambugu.

For more advanced understanding of these topics' students can also refer to these two textbooks.

7. Speech and Language Processing (3rd ed. draft) by Dan Jurafsky and James H. Martin
8. Deep Learning. Adaptive Computation and Machine Learning Series by I. Goodfellow, Y. Bengio, and A. Courville. London, England: MIT Press.

Assessments

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

Assessment	Percentage of Final Grade
Class Participation	10%
Two Class Quizzes	20%
Individual Assignment	20%
Team Project Presentation	40%
Total	100%

- **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.

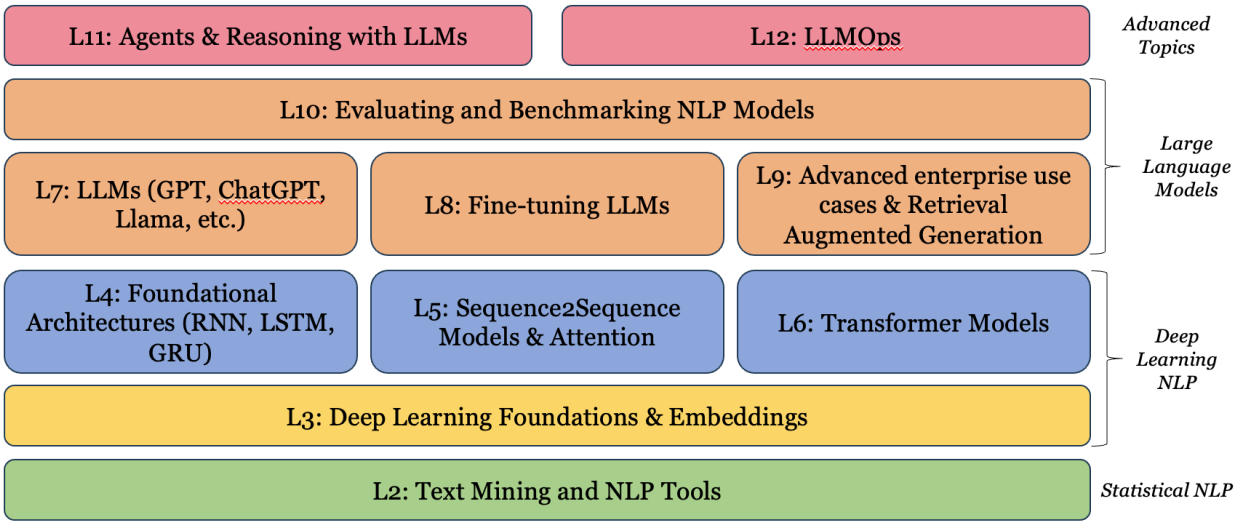
Course Schedule

Date	Theme/Topic	Learning Outcomes Addressed	Assignments/Quizzes
M: Aug 26	L1: Introduction to NLP: History, Evolution, and Importance	• Overview of the course, expectations, and logistics. • Understand NLP's definition, evolution, taxonomy, and its applications in Agentsdifferent sectors.	
W: Aug 28	L2: Text Mining and NLP Tools: Practical Techniques and application	• Acquire and apply text mining techniques, execute web scraping, perform sentiment analysis, and utilize TF-IDF for data analysis tasks.	(Ungraded) Homework Report Assignment-NLP-LLM-Resources-1
F: Aug 30	Recitation	• Exposure to PSC, NLP tools (NLTK), demos, and code walkthrough of material from L1&L2	(Ungraded) Quiz-L1&L2
M: Sep 2	No Class	• Labor Day Holiday	
W: Sep 4	L3: Deep Learning Foundations and Embeddings	• Master the use of neural networks in NLP, train models, understand embeddings, and assess learning through hands-on Tensor Playground and context-guessing activities	(Ungraded) Homework Programming Assignment-Text-Classification-2
F: Sep 6	Recitation	• Exposure to NLP frameworks (Keras, Spacy), demos and code walkthrough of material from L3.	(Ungraded) Quiz-L3
M: Sep 9	L4: Foundational Architectures: RNNs, LSTMs, and GRUs	• Demonstrate proficiency in RNN-based architectures (e.g., simple RNN, LSTM, GRU) and their NLP applications, including language detection, POS tagging, and NER through practical exercises and model analysis.	(Ungraded) Homework Programming Assignment-EmbeddingSpace-3

W: Sep 11	L5: Sequence to Sequence models & Attention	• Apply Sequence to Sequence models and Attention mechanisms to build and assess Machine Translation systems, Chatbots, and Text Classification solutions.	(Ungraded) Homework Programming Assignment-HateSpeech-4
F: Sep 13	Recitation	• Exposure to BERT and variants, demos and code walkthrough of material from L4&L5.	(Ungraded) Quiz-L4&L5
M: Sep 16	L6: Transformer models	• Analyze Transformer architectures and the foundational role they play in NLP tasks like spell checking, sentiment analysis, and inference, through hands-on use of BERT and its variants.	Quiz – 1
W: Sep 18	L7: Large Language Models - GPT, ChatGPT and other LLMs	• Examine and apply Large Language Models (LLMs) like GPT and Llama, exploring prompt engineering, and applications including semantic search, through practical exercises using OpenAI Playground and Llama.	(Ungraded) Homework Programming Assignment-Summarization-5
F: Sep 20	Recitation	• Exposure to GPT-4, Llama-3, and Mistral, demos and code walkthrough of material from L6&L7.	(Ungraded) Quiz-L6&L7
M: Sep 23	L8: Fine-tuning LLMs	• Implement and evaluate multiple fine-tuning methods for LLMs, applying these techniques to domain-specific tasks like sentiment and category classification of product reviews.	(Ungraded) Homework Programming Assignment-DebateGeneration-6
W: Sep 25	L9: Evaluating and Benchmarking NLP Models	• Master LLM evaluation using benchmarks like GLUE, SuperGLUE, etc., and assess model performance with metrics like perplexity, accuracy, BLEU scores ROUGE scores, etc., incorporating ethical considerations.	Quiz - 2
F: Sep 27	Recitation	• Exposure to model evaluation frameworks/tools and review material from L8 and L9.	Individual Programming Assignment Due (Ungraded) Quiz-L8&L9
M: Sep 30	L10: Advanced enterprise use	• Implement Retrieval Augmented Generation (RAG) for enterprise	(Ungraded) Homework Programming Assignment-

	cases & Retrieval Augmented Generation	applications, demonstrating its benefits through the development of a recommendation system.	Benchmarking-7
W: Oct 2	Guest Lecture (TBD)	• Insight into practical applications of LLMs.	
F: Oct 4	Recitation	• Exposure to RAG frameworks and tools and review material from L10.	Team Programming Interim Report Due (Ungraded) Quiz-L10
M: Oct 7	L11: Agents and Reasoning with LLMs	• Investigate and apply LLMs in agent-based reasoning across various domains, utilizing frameworks like Autogen and TaskWeaver, and evaluate through benchmarks and real-world agent applications.	(Ungraded) Homework Report/Programming Assignment-LLMinHC-10
W: Oct 9	L12: LLMOps	• Master Large Language Model Operations (LLMOps) focusing on ethical NLP practices, model deployment, and cost management for pre-training and inferencing.	Quiz – 3
F: Oct 11	Recitation/Final	Final Revision OR Team Programming Presentation & Code submission (TBD)	(Ungraded) Quiz-11&L12

Course Architecture



Course Topics

1. Week 1:
   - Lecture 1: Introduction to NLP: History, Evolution, and Importance
     - Topics
       1. Overview of Natural Language Processing
       2. What is Natural Language Processing, Understanding, and Generation
       3. Evolution of NLP
       4. NLP Taxonomy
       5. NLP Applications in Healthcare and Financial Services
     - Demos
       1. CORENLP.run
     - Book Readings
       1. [Required] Chapter 1 "Introduction to Natural Language Processing" from Real-World Natural Language Processing.
     - Paper Readings
       1. [Required] A Taxonomy of Natural Processing
       2. [Required] The History of NLP
       3. [Required] NLP Timeline
       4. [Optional] Natural language processing: state of the art, current trends, and challenges
       5. [Optional] Exploring the Landscape of Natural Language Processing Research
       6. [Optional] What is Natural Language Generation?
   - Lecture 2: Text Mining and NLP Tools: Practical Techniques and Applications
     - Topics
       1. Text mining pipeline
       2. Web scraping
       3. Sentiment analysis

4. TF-IDF
- Readings
    1. [Required] Chapter 2 “Your First NLP Application” from Real-World Natural Language Processing.
- Demos
    1. Web scraping
    2. Sentiment Analysis
    3. TF-IDF

- Recitation - Exposure to PSC, NLP tools (NLTK), demos, and code walkthrough of material from L1&L2
    - Demos
        1. Introduction to PSC
        2. Web Scraping Demo
        3. AllenNLP setup and walkthrough [Section 2.2.4 of RWNLP]
        4. Sentiment Analyzer model [Section 2.4.3 of RWNLP]
        5. Evaluation and Deployment [Sections 2.7 and 2.8 of RWNLP]
        6. Run code for Sentiment Analysis and TF-IDF

2. Week 2:
- Lecture 3: Deep Learning Foundations and Embeddings
    - Topics
        1. History of deep learning and embeddings
        2. Basics of neural networks and their relevance to NLP
        3. Simple neural networks and activation functions
        4. Training NLP models
        5. Word and document embeddings
        6. Activity – Guess the target and the context
        7. Word2Vec and Doc2Vec models
    - Demos
        1. Tensor Playground
        2. Stanford Sentiment Treebank
        3. Word Embeddings
        4. Activity: Guess-the-context & Fill-in-the-blank
    - Readings
        1. [Required] Chapter 3 “Word and Document Embeddings” from Real-World Natural Language Processing.
- Recitation - Exposure to NLP frameworks (Keras, Spacy), demos and code walkthrough of material from L3.
    - Demos
        1. Introduction to Keras
        2. Introduction to Spacy
        3. Skip-Gram on AllenNLP (Section 3.4.5 of RWNLP]
        4. Gensim and Doc2Vec [Section 3.7 of RWNLP]
        5. Visualize GLoVe embeddings [Section 3.8 of RWNLP]
        6. Building AllenNLP training Pipelines [Section 4.4/4.5 of RWNLP]

3. Week 3:

- Lecture 4: Foundational Architectures: RNNs, LSTMs, GRUs, and Beyond
    - Topics
        1. Introduction to RNNs, LSTMs, and GRUs
        2. Applications, advantages, and disadvantages of RNNs, LSTMs, GRUs
        3. Applications: Language detection
        4. Sequential labeling & POS tagging
        5. Multi-layered and bi-directional RNNs
        6. Applications: Named Entity Recognition (NER)
        7. RNN-based Language model
    - Demos
        1. Activity – Memory match with RNN family
        2. Language detection
        3. Named Entity Recognition
    - Readings
        1. [Required] Chapter 4 "Sentence classification" from Real-World Natural Language Processing.
        2. [Required] Chapter 5 "Sequential Labeling and language modeling" from Real-World Natural Language Processing.
- Lecture 5: Sequence to Sequence models & Attention
    - Topics
        1. Sequence2Sequence models
        2. Encoders-Decoders
        3. Greedy and Beam search decoding
        4. Self-Attention
        5. Encoder-Decoder attention
        6. Applications: Machine Translation
        7. Applications: Chatbots
        8. Applications: Text Classification
    - Demos
        1. Machine Translation
        2. Building a chatbot
    - Book Readings
        1. [Required] Chapter 6 "Sequence-to-sequence Models" from Real-World Natural Language Processing.
        2. [Optional] Chapter 7 "Convolutional Neural Networks" from Real-World Natural Language Processing.
        3.
- Recitation - Exposure to BERT and variants, demos and code walkthrough of material from L4&L5.
    - Demos
        1. Language Detection [Section 4.6 of RWNLP]
        2. POS Tagger [Section 5.2 of RWNLP]
        3. NER [Section 5.4 of RWNLP]
        4. Language Model [Section 5.6.2 of RWNLP]
        5. FairSeq
        6. Machine Translation with Transformer

7. Building a chatbot
8. Spell-Checker
9. Sentiment Analysis with BERT
10. Natural language inference with BERT

4. Week 4:
   - Lecture 6: Transformer
     - Topics
       1. Types of Attention
       2. Encoder-decoder attention
       3. Transformer model
       4. Transformer-based language models
       5. Language Models
       6. Autoencoding and autoregressive language Models
       7. Transfer Learning
       8. BERT and its variants (e.g., FinBert, BioBert, DistilBert)
       9. Application: Spell-Checker
     - Demos
       1. Machine Translation with Transformer
       2. Spell-Checker
       3. Sentiment Analysis with BERT
       4. Natural language inference with BERT
     - Book Readings
       1. [Required] Chapter 8 "Attention and Transformer" from Real-World Natural Language Processing.
       2. [Required] Chapter 9 "Transfer Learning with pretrained language models" from Real-World Natural Language Processing.
   - Lecture 7: Large Language Models – GPT, ChatGPT, Llama
     - Topics
       1. Large Language Models.
       2. GPT and ChatGPT.
       3. Prompt engineering
       4. Taxonomy of prompting
       5. Batch prompting, prompt chaining, Chain-of-thought prompting
       6. Zero-shot and few-shot learning.
       7. Other LLMs - LLaMa, PaLM etc.
       8. Domain-specific LLMs
       9. Applications of LLMs
       10. Semantic Search
     - Book Readings
       1. [Required] Chapter 1 "Overview of Large Language Models" Quick Start Guide to Large Language Models
       2. [Required] Chapter 2 "Semantic Search with LLMs" Quick Start Guide to Large Language Models

3. [Required] Chapter 3 “First steps with Prompt Engineering” from “Transfer Learning with pretrained language models” from Quick Start Guide to Large Language Models

- Recitation - Exposure to GPT-4, Llama-3, and Mistral, demos and code walkthrough of material from L6&L7.
    - Demos
        1. ChatGPT, Gemini, Perplexity.ai
        2. Amazon Review Sentiment Classification
        3. Amazon Review Category Classification
    -

5. Week 5:
    - Lecture 8: Fine-tuning techniques LLMs
        - Topics
            1. Fine tuning with closed-source pre-trained models (e.g., OpenAI)
            2. Fine-tuning techniques – supervised fine-tuning, reinforcement learning from human feedback
            3. Qualitative evaluation of LLMs
            4. Applications: Amazon Review Sentiment Classification
            5. Applications: Amazon Review Category Classification
        - Book Readings
            1. [Required] Chapter 4 “Optimized LLMs with Customized Fine-Tuning” from “Transfer Learning with pretrained language models” from Quick Start Guide to Large Language Models
            2. [Required] Chapter 5 “Advanced Prompt Engineering” from “Transfer Learning with pretrained language models” from Quick Start Guide to Large Language Models
    - Lecture 9: LLM Evaluation and benchmarking
        - Topics
            1. Importance and challenges of model evaluation and benchmarking
            2. Standard metrics for evaluation – perplexity, accuracy, F1-score, ROUGE score, BLEU score, METEOR score, etc.
            3. Standard benchmarks and dataset – GLUE, SuperGLUE, MMLU, BIG-Bench, LAMA, HELLA-Swag, TruthfulQA, AI2 Reasoning Challenge, HELM, etc.
            4. Ethical considerations in LLM Evaluation
        - Demos
            1. DeepChecks testing and monitoring (https://github.com/deepchecks/deepchecks)
            2. LLM Benchmarker Suite (https://github.com/TheoremOne/llm-benchmarker-suite)
            3. HELM Classic (https://crfm.stanford.edu/helm/classic/latest/)
            4. HELM Lite (https://crfm.stanford.edu/helm/lite/latest/)
    - Recitation - Exposure to model evaluation frameworks/tools and review material from L8 and L9.
        - Demos
            1. HELM
            2. Zeno

6. Week 6:
    - Guest Lecture -1: Practical Applications of LLM in Financial Services
        - Overview of NLX and LLM applications in Financial Services

        - Pre-training Domain-Specific Models
- Lecture 10: Advanced enterprise use cases & Retrieval Augmented Generation
    - Topics
        1. Retrieval Augmented Generation (RAG)
        2. Benefits of RAG in enterprise applications
        3. Real-world examples of RAG in action
        4. Applications: Building a Recommendation system
    - Book Readings
        1. [Required] Chapter 6 "Customizing Embeddings and Model Architectures" from Quick Start Guide to Large Language Models
- Recitation - Exposure to RAG frameworks and tools and review material from L10.
    - Demos
        1. RAG frameworks

7. Week 7:
- Lecture 11: Agents and Reasoning with LLMs
    - LLMs in temporal reasoning, planning, commonsense, causal, visual, audio, multi-modal, memory, research, and arithmetic
    - Benchmarking reasoning LLMs
    - Agent frameworks like Autogen, TaskWeaver
    - Applications of agents for simulating human behavior, urban contexts, social impact, engineering, natural science, game playing etc.
    - Required Readings
        1. [Required] A Survey of Reasoning with Foundational Models
        2. [Required] Exploring LLM-based Intelligent Agents: Definitions, Methods, and Prospects
        3. https://github.com/WooooDyy/LLM-Agent-Paper-List
    - Demos
        1. Autogen (https://microsoft.github.io/autogen/)
        2. AutoGen with Ollama/LiteLLM (https://www.youtube.com/watch?v=y7wMTwJN7rA)
        3. Taskweaver (https://www.youtube.com/watch?v=y7wMTwJN7rA)
        4. Interactive Simulacra of human behavior (https://reverie.herokuapp.com/arXiv_Demo/)
        5. Multi-agent Hide and Seek (https://www.youtube.com/watch?v=kopoLzvh5jY)
        6. AutoGPT (https://github.com/Significant-Gravitas/AutoGPT)
- Lecture 12: Large Language Model Operations (LLMOps)
    - Ethical considerations in NLP and potential risks
    - Best practices in developing NLP applications – batching instances, tokenization, avoiding overfitting, dealing with imbalanced datasets, and hyperparameter tuning.
    - Deploying and serving NLP applications – architecture, deployment, operations, observability, and visualization
    - Deploying closed-source and open-source LLMs into production
    - LLM pre-training, fine-tuning and inferencing cost considerations

- Book Readings
    3. [Required] Chapter 10 “Best Practices in Developing NLP Applications” from Real-World Natural Language Processing.
    4. [Required] Chapter 11 “Deploying and Serving NLP Applications” from Real-World Natural Language Processing.
    5. [Required] Chapter 9 “Moving LLMs into production” Quick Start Guide to Large Language Models
- Demos
    1. Case Study – Distilling Our Anime Genre Predictor
    2. Hugging Face deployment

- Recitation
    - Course Revision
    - Demos

This course is designed for graduate-level students who have a programming and analytics background, exposed to systems thinking and are proficient in Python programming.