Lahore University of Management Sciences

CS 438/CS 5302/EE 416/EE 519 - Speech and Language Processing with Generative AI

Fall 2023-24

Course description
Generative AI stands at the cutting edge of today's artificial intelligence landscape, ushering in a new paradigm where machines not only understand intricate data patterns but also autonomously produce them. This in-depth course ventures into the fascinating world of Generative AI, cultivating a deep understanding of its potential to adeptly create, communicate, and innovate across diverse data forms. Students will gain hands-on experience with some of today's most renowned models, adapting them to unique use-cases while engaging with a vast array of topics—from foundational theories and principles to design, hands-on implementation, and thorough analysis of these systems. By the course's end, participants will be equipped to transition into the industry with tangible skills and a robust portfolio, contribute meaningfully to academic discourse by augmenting existing research or pioneering novel concepts, or embark on personal projects with an enhanced perspective and expertise.

Course distribution	
Elective	This is a Graduate Level CS elective course to be cross-listed as an undergraduate elective course.
Open for Student Category	Juniors, seniors, and graduates.
Close for Student Category	Please see the prerequisites below.

Course prerequisites
• All students must have taken CS535/EE514 (Machine Learning)

Course Offering Details				
Credit Hours	3 hours			
Lecture(s)	Nbr of lec(s) per week	2	Duration	75 minutes
Recitation/Lab (per week)	Nbr of lec(s) per week		Duration	
Tutorial (per week)	Nbr of lec(s) per week	1 (optional)	Duration	50 minutes

Instructor	Agha Ali Raza
Room No.	SBASSE 9-G49A
Office Hours	TBA
Email	agha.ali.raza@lums.edu.pk
Telephone	8565
Secretary/TA	TBA
TA Office Hours	TBA
Course URL (if any)	None

Course Teaching Methodology (Please mention the following details in plain text)

- **Lectures:** In-person.
- **TA Sessions:** TAs will conduct asynchronous and synchronous sessions (in-person and online) to cover tutorials related to assignments.
- **Exams:** Exams will be conducted in person in pre-scheduled sessions.
- **Quizzes:** Quizzes will be conducted during announced class timings.
- **Class discussions:** There will be a slack channel for all discussions (general, assignments, quizzes, etc.)

PROGRAM EDUCATIONAL OBJECTIVES (PEOs)

PEO-01	Demonstrate excellence in the profession through in-depth knowledge and skills in the field of Computing.
PEO-02	Engage in continuous professional development and exhibit a quest for learning.
PEO-03	Show professional integrity and commitment to societal responsibilities.

Course Objectives

The goal of this course is to

- Be able to utilize foundation models and tune them in order to achieve a well-defined goal
- Understand the theory of the ideas behind, and the applications of modern generative systems including Large Language Models
- Get hands-on experience with the creation of AI-powered applications
- Gain a firm grip on expertly critiquing and analyzing the entire pipeline of a system

COURSE LEARNING OUTCOMES (CLOs)

By the end of the course, students should be able to:

CLO1:	Understand the foundations and core ideas of modern Machine Learning architectures and models
CLO2:	Appreciate the scope of datasets and variety of training mechanisms
CLO3:	Learn core ideas in creating applications using these models, including use-cases and practical details
CLO4:	Understand the evolution of foundation models in terms of techniques, scale, working mechanisms etc.

CLO	CLO Statement	Bloom's Cognitive Level	PLOs/Graduate Attributes (Seoul Accord)
CLO1			
CLO2			
CLO3			
CLO4			

Grading Breakup and Policy

Assessment	Weight (%)	Related CLOs	ACM Recommended Disposition
Assignments	25%		
Quizzes	20%		
Paper Presentations	20%		
Project	25%		
Final Exam	10%		

Examination detail

Midterm Exam	Yes/No:	No
	Duration:	
	Exam Specifications:	
Final Exam	Yes/No:	Yes
	Duration:	2.5 – 3 hours
	Exam Specifications:	In-person exam

SSE Council on Equity and Belonging

In addition to LUMS resources, SSE's **Council on Belonging and Equity** is committed to devising ways to provide a safe, inclusive and respectful learning, living, and working environment for students, faculty and staff. To seek counsel related to any issues, please feel free to approach either a member of the council or email at cbe.sse@lums.edu.pk.

Mental Health Support at LUMS

For matters relating to counseling, kindly email student.counselling@lums.edu.pk, or visit https://osa.lums.edu.pk/content/student-counselling-office for more information. You are welcome to write to me or speak to me if you find that your mental health is impacting your ability to participate in the course. However, should you choose not to do so, please contact the Counseling Unit and speak to a counselor or speak to the OSA team and ask them to write to me so that any necessary accommodations can be made.

Harassment Policy

SSE, LUMS and particularly this class, is a harassment free zone. Harassment of any kind is unacceptable, whether it be sexual harassment, online harassment, bullying, coercion, stalking, verbal or physical abuse of any kind. Harassment is a very broad term; it includes both direct and indirect behavior, it may be physical or psychological in nature, it may be perpetrated online or offline, on campus and off campus. It may be one offense, or it may comprise of several incidents which together amount to sexual harassment. It may include overt requests for sexual favors but can also constitute verbal or written communication of a loaded nature. Further details of what may constitute harassment may be found in the LUMS Sexual Harassment Policy, which is available as part of the university code of conduct.

LUMS has a Sexual Harassment Policy and a Sexual Harassment Inquiry Committee (SHIC). Any member of the LUMS community can file a formal or informal complaint with the SHIC. If you are unsure about the process of filing a complaint, wish to discuss your options or have any questions, concerns, or complaints, please write to the Office of Accessibility and Inclusion (OAI, oai@lums.edu.pk) and SHIC (shic@lums.edu.pk) —both of them exist to help and support you and they will do their best to assist you in whatever way they can. You can find more details regarding the LUMS sexual harassment policy here.

To file a complaint, please write to harassment@lums.edu.pk.

Rights and Code of Conduct for Online Teaching

A misuse of online modes of communication is unacceptable. TAs and faculty will seek consent before the recording of live online lectures or tutorials. Please ensure if you do not wish to be recorded during a session to inform the faculty member in a timely manner. Please also ensure that you prioritize formal means of communication (email, LMS) over informal means to communicate with course staff.

Course overview

Week	**Topics**	**Recommended Readings**	**Related CLOs**	**ACM Comp Knowledge Landscape**
1.	**Course Overview** ● A history for Machine Learning ● What is NLP/NLU? ○ The Boom for Language Technologies ○ Examples of Applications ● What is Generative AI? ○ Generative AI for language ○ Generative AI for speech ○ Generative AI for vision ● Opportunities of ML ○ ML for social good, ML for Development (ML4D), Language Technologies for Development (LT4D) **Basics of Natural Language Processing** ● Natural language (and human speech) ● Subdomains in NLP and their applications ○ Phonetics and phonology ○ Morphology ○ Syntax ○ Semantics ○ Discourse and pragmatics ● Introduction to Python and the Natural Language Toolkit (NLTK) ○ English and Urdu Corpus processing ● Regular Expressions ● Normalization and collation; Surface form and deep structure; types and tokens; root, lexeme, lemma ● Word formation processes: Inflection, derivation, compounding, cliticization, reduplication			

	● Word and Sentence tokenization
	● Stem, stemming, Information Retrieval
	● Morphology and Morphological Processing
	● Language, script and style
	● String similarity and distance
	● Vector similarity and distance measures
	    ○ Euclidean distance
	    ○ Manhattan distance
	    ○ Chebyshev distance
	    ○ Minkowski distance
	    ○ Cosine similarity
	● Other string similarity and distance measures
	    ○ Jaccard Similarity
	    ○ Jaro similarity
	    ○ Jaro-Winkler similarity
	    ○ Edit distance
	        ■ Levenshtein distance
	        ■ Damerau–Levenshtein distance
	        ■ Longest common subsequence (LCS)
	        ■ Hamming distance
	● o Phonetic similarity			

2.	**Speech and Language Processing**
	● Notion of Sequence Modeling tasks
	    ○ Text and Speech
	● Recurrent Neural Networks
	    ○ Overall Architecture
	    ○ The Hidden State
	    ○ "Unrolling" a unit
	● LSTMs
	    ○ Changes to the Architecture
	    ○ Storing the "memory" in a cell
	● Machine Translation and Embeddings
	    ○ Setup for a Seq2Seq problem
	    ○ Tokenization
	    ○ Embeddings as vector representations
	    ○ Encoder-Decoder framework
	    ○ Information bottleneck: passing on only one hidden state
	    ○ Improvements
	        ■ Passing all the hidden states
	        ■ The Attention Mechanism
		NLP Review Levels of analysis: phonetics, phonology, morphology, syntax, semantics, pragmatics, discourse
		
		All terminology of NLP
		
		From my NLP outline.
		2 lectures
		
		The Unreasonable Effectiveness of RNNs
		
		Visualizing A Neural Machine Translation Model		

3.	**Attention and Transformers**
	● The Attention Mechanism in Machine Translation
	● Self-Attention
	    ○ Dot Product Attention
	    ○ Contextualized Token Embeddings
	● The Transformer and its advantages
	    ○ Highly Parallelizable
	    ○ Contextualized Embeddings vs. Regular Embeddings
	    ○ Long-Term Dependencies
	● Transformer Architecture in a nutshell
	    ○ Attention Is All You Need
	    ○ The Encoder
	    ○ The Decoder and Masked Attention
	    ○ Query, Key, Value from tokens
	● The Transformer in equations
	    ○ Positional Embeddings
		Attention Is All You Need
		
		Jay Alammar's The Illustrated Transformer
		
		Some Intuition on Attention and the Transformer
		
		The Annotated Transformer		

	○ Projections to QKV ○ Self Attention as Dot Product Attention ○ Role of Feedforward Layers ○ Multi-Headed Self Attention	Transformers from Scratch		
4.	**Pre-Training and Transfer Learning** ● Pre-training objectives vs. downstream tasks ○ Masked Language Modeling as a pre-training objective ● Transformer Case Studies ○ BERT: Bidirectional Encoder Representations from Transformers ■ Encoder-based Transformer ■ Generating Embeddings ■ Scaling up BERT: S to XL ○ T5: Text-to-Text Transfer Transformer ■ Encoder-Decoder Transformer ■ Translation capabilities ● Leveraging pre-trained models ○ Fine-tuning models for downstream tasks ● Prompt Engineering ○ Simple Prompts ○ Chain-Of-Thought ○ In-Context Learning	The Illustrated BERT The State of Transfer Learning in NLP Universal Language Model Fine-tuning for Text Classification Exploring Transfer Learning with T5 BERT: Bidirectional Encoder Representations from Transformers T5: Text-to-Text Transfer Transformer		
5.	**Instruction-Tuned Models** ● Generative Pretrained Transformer (GPT) ○ GPT-3 Case Study ○ Training Cycle ■ Pretraining, Supervised Fine-Tuning, RLHF ■ State of GPT ● Instruction Tuning ○ Issue with Alignment ○ Relation to Pre-training and Fine-tuning ○ Case Study: InstructGPT, Codex ● Proprietary vs. Open-Source LLMs ○ Drawbacks of reliance on Proprietary Models ○ The boom with Open-Source LLMs ● Alpaca and LLaMA ○ Mining datasets ○ Scale of the models ○ Comparisons to proprietary counterparts	What We Know About LLMs (Primer) State of GPT ChatGPT InstructGPT Language Models are Few-Shot Learners (GPT-3) LLama 2: Open Foundation and Fine-Tuned Chat Models Jessu Mu - Prompting		
6.	**Case Studies for Specialized LLMs** ● Gorilla ○ Finetuning a model to use APIs ○ Mitigating Hallucinations ○ Document Retriever to make updates ● Orca ○ Imitation learning ○ Imitating the reasoning process, not the style of LLMs ○ Finetuning on explanation traces ● Goat ○ Improving mathematical abilities of LLMs ○ Tokenization of numbers ○ Simplistic synthetic datasets ● Evaluation of LLMs ○ Perplexity ○ BLEU ● Hallucinations in LLMs	Gorilla: Large Language Model Connected with Massive APIs Orca: Progressive Learning from Complex Explanation Traces of GPT-4 Goat: Fine-tuned LLaMA Outperforms GPT-4 on Arithmetic Tasks		

7.	**Fine-Tuning Paradigms I**
	● Conventional Fine-Tuning
	  ○ Making all parameters trainable
	  ○ Freezing parameters
	● In-Context Learning
	  ○ Mimicking Gradient Descent
	  ○ Few-Shot vs. Zero-Shot performance
	● Parameter Efficient Fine-Tuning (PEFT)
	  ○ Limitations of compute
	  ○ Scaling up models
	  ○ PEFT vs. Conventional Fine-Tuning
	● Low Rank Adaptation (LoRA)
	  ○ Injecting randomly initialized parameters
	  ○ Feasibility on lower-tier machines
	● Quantization
	  ○ Varying the precision of parameters
	  ○ 32-bit vs. 16-bit vs. 8-bit vs. 4-bit
	  ○ Case Study: fp16 vs. bfloat16
		LoRA: Low Rank Adaptation of Large Language Models
		
		Finetuning Large Language Models
		
		Explaining the Key Concepts Behind LoRA
		
		Parameter Efficient Fine-Tuning (blog)
		
		Model Training Anatomy		

8.	**Fine-Tuning Paradigms II**
	● Quantized Low Rank Adaptation (QLoRA)
	  ○ An amalgamation of different techniques
	  ○ NF4, Nested Quantization, Paging Optimizers
	  ○ Fine-Tuning 20B LLMs on free-tier cloud instances
	  ○ Inference Scaling Laws
	
	● Data Quality Influences on Model Performance
	  ○ Mining methods
	    ■ Bonafide vs. Synthetic Data
	  ○ The False Promise of Imitating Proprietary LLMs
	  ○ Less Is More for Alignment (LIMA)
	  ○ Textbooks Are All You Need
		QLoRA: Efficient Finetuning of Quantized LLMs
		
		QLoRA Is All You Need
		
		The False Promise of Imitating Proprietary LLMs
		
		Less Is More for Alignment (LIMA)
		
		Textbooks Are All You Need		

9.	**Systems with NLP**
	● Designing a System
	  ○ Using LLMs out of the box vs. Fine-Tuning
	  ○ Creating a Pipeline
	  ○ Deployment options
	● Information Retrieval Mechanisms
	  ○ Utilization of External Knowledge
	  ○ Vector Stores and Vector Databases
	  ○ Semantic Search
	  ○ Retrieval Augmented Generation (RAG)
	  ○ Mitigating the Hallucination Issue
	  ○ Case Study: Retrofit Attribution Using Research and Revision (RARR) for Fact-checking with LLMs
		Building RAG-based LLM Applications for Production
		
		Explaining Vector Databases in 3 Levels of Difficulty
		
		Patterns for Building LLM-based Systems and Products
		
		Using BERT pre-trained Embeddings directly for Semantic Search		

10.	**Multilingual NLP**
	● Challenges in training on languages outside of English
	  ○ Lack of data
	  ○ Differences between languages
	  ○ Skewed language distributions in large datasets
	● Multilingual BERT (mBERT)
	  ○ Dataset and mining techniques
	  ○ Training mechanism
	● Case Study: Multilingual Representations for Indian Languages (MuRIL)
	  ○ Dataset aggregation
	  ○ Training mechanism
	  ○ Improvements over mBERT
		How Multilingual is Multilingual BERT?
		
		MuRIL: Multilingual Representations for Indian Languages
		
		State of Multilingual AI
		
		Why you should do NLP beyond English
		
		A Primer on Pretrained		

		Multilingual Language Models		
11.	**Multimodal Models** ● What are Multimodal Models? ○ Vision and Language ○ Language and Speech ○ Examples of tasks ■ Image Captioning ■ VQA ■ Zero-shot Image Classification ● Background Concepts ○ Multimodal Fusion ○ Cross-modal Attention Mechanisms ○ Deep Learning for Computer Vision ● Multimodal Models for Vision and Language ○ Vision-Language Pretraining ○ Contrastive Learning ○ Introduction to Contrastive Language-Image Pretraining (CLIP)	Douwe Kiela - Multimodal Deep Learning Multimodal Machine Learning CVPR 2022 Tutorial (series) How Multimodal Models are leading the way Fundamentals of Multimodal Representation Learning		
12.	**Multimodal Models II** ● Contrastive Language-Image Pretraining (CLIP) ○ Architecture ○ Training Mechanism ● LLaVA: Large Language and Vision Assistant ● Applications ○ Visual Question Answering ○ Optical Character Recognition ○ Action Recognition from Video ○ Object Classification	Foundation Models - CLIP Learning Transferable Visual Models from Natural Language Supervision The Annotated CLIP (Part 1, Part 2) LLaVA Webpage Visual Instruction Tuning		
13.	**Explainability** ● The importance of Explainable AI (XAI) ○ Motivation and the need for explainability ● Intrinsic vs. Post-hoc Explainability ○ Intrinsic Explainability in model design ○ Post-hoc techniques for existing models ● Visualization techniques ○ Heatmaps for feature importances ○ Attention maps to understand relations	What is Explainable AI? Introduction to Explainable AI (ML Tech Talks)		
14.	**Ethics** ● Cases of Misuse ○ Misinformation and its impact ○ Creation and detection of fake news ○ Legal and Ethical implications ○ Media literacy ● Bias and Fairness ○ Algorithmic bias ○ Building systems with ethical considerations ● Privacy and Surveillance ○ Implications for Cybersecurity ○ GDPR and the "right to explanation" ○ Countermeasures	Word Embeddings, Bias in ML, Why You Don't Like Math & Why AI Needs You 21 Definitions of Fairness How Algorithms Can Learn to Discredit "the Media" The Problem with Metrics		
	Other topics – to be covered if we have time			

Textbook(s)/Supplementary Readings
Speech and Language Processing by Jurafsky and Martin, 3rd edition

Course policies
Use of electronic devices (e.g., mobile phones and laptops) in the class is strictly forbidden. A violation could result in deduction of marks and other strict penalties **Late arrival:** You may not be allowed in the class 10 minutes after the start time **Plagiarism:** All work MUST be done independently. In certain assignments students will be allowed to have discussions with peers, in which case they must mention the name and roll number of the student with whom the discussion took place and the nature of the discussion. Even in those assignments, all implementations need to be done independently. Any plagiarism or cheating of work from others or the internet will be immediately referred to the DC. If you are confused about what constitutes plagiarism, it is YOUR responsibility to consult with the instructor or the TA in a timely manner. No "after the fact" negotiations will be possible. ● Submitting someone else's assignment as your own "by mistake" would count as plagiarism. If this indeed happens accidentally, please let us know immediately (within minutes) along with an explanation and do not wait until we find it out on our own. In the latter case, it would be considered plagiarism. **Quizzes:** Quizzes will be unannounced. We will be following an n-x (x=2) policy for the quizzes. <u>There is no makeup for a missed quiz</u>. If you have missed up to x quizzes, you will be covered only using the n-x policy (even if you have an approved petition with the OSA). If you have missed more than x quizzes, then you would be awarded the average marks (across all the quizzes that you attempted) for each missed quiz, provided that your case has been approved by the Office of Student Affairs. **Non-uniform weightage:** All subcomponents (e.g., quizzes, assignments) may not carry the same weight. These weights may not be announced prior to the submission of the components and will be determined by the course instructor based on factors including (but not limited to) the length, difficulty level, amount of help available, etc. for each subcomponent. **Programming:** Strong programming skills are expected for this course. Please keep in mind that this is a programming intensive course, and you will be spending a lot of time designing and coding up your solutions. **Assignments:** There is negative marking for skipped assignments and there is no n-x policy for assignments. Assignments are a basic building block of this course, and it will be ensured that students, who pass the course, have significant hands-on experience. ● You will be awarded 0 marks or investigated for plagiarism for submitting incorrect/corrupted files and/or older assignments. We will not accept resubmissions in these cases even if the system date shows that the file was not modified after the deadline. ● You are allowed 5 grace days for the entire semester. No late submission of assignments is allowed after your grace days have expired. We do not have any deduction policy for late submissions in addition to the grace days. All grace days must be utilized before the start of the dead week and any remaining grace days will expire as soon as the dead week begins. ● Please do not wait until the last moment to submit assignments and other components. Any requests to accommodate late submissions due to last minute issues (submission of partial or incorrect files, assignment server down-time, internet and power failures, personal problems, etc.) would not be accommodated.

Appendix A
Bloom's Taxonomy

BLOOM's TAXONOMY*	
1 - Remember 2 - Understand 3 - Apply 4 - Analyze 5 - Evaluate 6 - Create	● Recall facts and basic concepts ● Explain ideas or concepts ● Use information in new situations ● Draw connection among ideas ● Justify a stand or decision ● Produce new or original work

https://cft.vanderbilt.edu/guides-sub-pages/blooms-taxonomy/

Appendix B

ACM Dispositions Table - I

ACM Dispositions			
Element	**Elaboration**	**Element**	**Elaboration**
D1 Adaptable:	Flexible; agile, adjust in response to change	D7 Professional:	Professionalism, discretion, ethical, astute
D2 Collaborative:	Team player; willing to work with others	D8 Purpose-driven:	Goal driven, achieve goals, business acumen
D3 Inventive:	Exploratory; Look beyond simple solutions	D9 Responsible:	Use judgment, discretion, act appropriately
D4 Meticulous:	Attentive to detail; thoroughness, accurate	D10 Responsive:	Respectful; react quickly and positively
D5 Passionate:	Conviction, strong commitment, compelling	D11 Self-directed:	Self-motivated, determination, independent
D6 Proactive:	With initiative, self-starter, independent		

ACM Dispositions Table - II

Class Assessments and Proposed Dispositions												
Assessment Type	**D1 Adaptable**	**D2 Collaborative**	**D3 Inventive**	**D4 Meticulous**	**D5 Passionate**	**D6 Proactive**	**D7 Professional**	**D8 Purpose-driven**	**D9 Responsible**	**D10 Responsive**	**D11 Self-directed**	**Included**
Quiz				✓			✓		✓			Yes
Assignment-Individual			✓	✓			✓		✓			Yes
Assignment-Group		✓	✓	✓			✓		✓	✓		Yes
Project-Individual	✓		✓	✓	✓	✓	✓	✓	✓		✓	Yes
Project-Group	✓	✓	✓	✓	✓	✓	✓	✓	✓			Yes
Presentation - Individual				✓			✓		✓	✓	✓	Yes
Presentation - Group		✓		✓			✓		✓	✓		Yes
Labs-Individual			✓	✓			✓		✓			Yes
Labs- Group		✓	✓	✓			✓		✓	✓		Yes
Exams				✓			✓		✓			Yes
Included	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	

Appendix C
ACM Computing Knowledge Landscape Table

ACM Computing Knowledge Landscape (CK)			
1. **Users and Organizations**	CK1.1: Social Issues and Professional Practice CK1.2: Security Policy and Management CK1.3: IS Management and Leadership CK1.4: Enterprise Architecture CK1.5: Project Management CK1.6: User Experience Design	4. **Software Development**	CK4.1: Software Quality, Verification and Validation CK4.2: Software Process CK4.3: Software Modeling and Analysis CK4.4: Software Design CK4.5: Platform-Based Development
2. **Systems Modeling**	CK2.1: Security Issues and Principles CK2.2: Systems Analysis & Design CK2.3: Requirements Analysis and Specification CK2.4: Data and Information Management	5. **Software Fundamentals**	CK5.1: Graphics and Visualization CK5.2: Operating Systems CK5.3: Data Structures, Algorithms and Complexity CK5.4: Programming Languages CK5.5: Programming Fundamentals CK5.6: Computing Systems Fundamentals
3. **Systems Architecture and Infrastructure**	CK3.1: Virtual Systems and Services CK3.2: Intelligent Systems (AI) CK3.3: Internet of Things CK3.4: Parallel and Distributed Computing CK3.5: Computer Networks	6. **Hardware**	CK6.1: Architecture and Organization CK6.2: Digital Design CK6.3: Circuits and Electronics CK6.4: Signal Processing