Generative AI:Security, Ethics and Governance

CS 4XX AM

Blueprint Term – Spring 2025

“With great power comes great responsibility” - *Uncle Ben, Spider-Man*

“We talk a lot about building benevolent technology. Our technology reflects our values.” *-Fei-Fei Li*

“The development of full artificial intelligence could spell the end of the human race” - *Stephen Hawking*

Credit Hours and Pre-requisites

Credit Hours: 3+0

Prerequisite Courses:) CS 351 Artificial Intelligence, 2) CS 335 Introduction to Large Language Model, 3) CS 316 Introduction to Deep Learning, 4) CS 458 Natural Language Processing, 5) EE 452 Computer Vision

Taking any one of these courses will fulfill the prerequisite requirement

Instructor Information

Instructor: Omema Rizvi
Title: Research Assistant (Computer Science)
Office Location: C100
Email: omema.rizvi@sse.habib.edu.pk
Office Hours: TBA

Course Description

The course covers Generative AI and LLM fundamentals, security challenges like prompt injection attacks and adversarial examples, cloud-specific AI security for AWS, GCP, and Azure, AI governance frameworks, and ethical considerations in AI development. It features hands-on assignments and a final project focused on implementing secure LLM solutions.

Course Aims

To equip students with a comprehensive understanding of security challenges, governance frameworks, and ethical considerations in Generative AI. Students will gain hands-on experience in implementing security measures for LLMs with cloud providers, analyzing regulatory impacts, and addressing ethical dilemmas in AI development and deployment.

Course Learning Outcomes (CLOs)

CLO 1 - Explain the fundamentals of Generative AI, LLMs, and their architectures.

CLO 2 - Analyze and implement security measures for LLMs.

CLO 3 - Evaluate the impact of AI regulations and governance frameworks.

CLO 4 - Apply ethical principles in AI development and deployment.

CLO 5 - Implement secure LLM solutions in various cloud environments and open-source platforms.

Mode of Instruction

The class will be conducted once a week for 150 minutes in live online class mode. Students can expect to work for at least 6-8 hours per week outside of the scheduled class for this course.

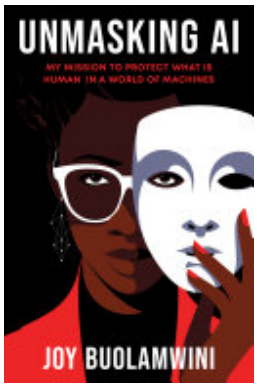
Engagement & Participation Rules

Active participation in discussions, completion of hands-on exercises, contributions to projects, and timely submission of assignments are expected.

Required Texts and Materials

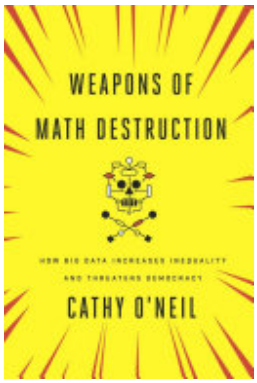
Responsible AI in the Enterprise
ISBN: 9781803249667

Authors: Adnan Masood, Heather Dawe
Publisher: Packt Publishing Ltd
Publication Date: 2023-07-31

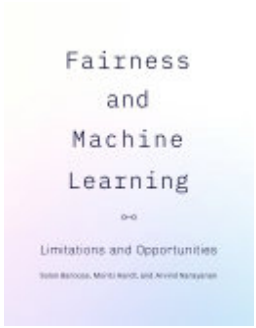


Unmasking AI
ISBN: 9780593241837
Authors: Joy Buolamwini
Publisher: Random House
Publication Date: 2023-10-31

Optional Materials



Weapons of Math Destruction
ISBN: 9780553418811
Authors: Cathy O'Neil
Publisher: Crown Publishing Group (NY)
Publication Date: 2016-01-01



Fairness and Machine Learning
ISBN: 9780262048613
Authors: Solon Barocas, Moritz Hardt, Arvind Narayanan
Publisher: MIT Press
Publication Date: 2023-12-19

Assessments

The final project will involve developing a comprehensive security and governance framework for a hypothetical LLM-based application. Details will be provided mid-semester.

	Frequency	Grade
Assignments	12	75%
Quizzes	10	10%
Final Project	01	15%

Grading Scale

Letter Grade	GPA Points	Percentage
A+	4.00	[95-100]
A	4.00	[90-95)
A-	3.67	[85-90)
B+	3.33	[80-85)
B	3.00	[75-80)
B-	2.67	[70-75)
C+	2.33	[67-70)
C	2.00	[63-67)
C-	1.67	[60-63)
F	0.00	[0, 60)

Note: [a, b) is a range of numbers from a to b where a is included in the range and b is not.

Late Submission Policy

Late submissions will incur a 10% grade reduction for every 24 hours of delay, unless prior arrangements have been made with the instructor.

Week-Wise Schedule (Tentative)

Spring 2024 Weekly Schedule*

Week	Sessions	Readings	Assessments

			and Due Date
Week - 1 January 8 – 12, 2024	Introduction to Generative AI Large Language Models (LLMs)	”Language Models are Few-Shot Learners” (Brown et al., 2020) On the Opportunities and Risks of Foundation Models” (Bommasani et al., 2021)	Assignment 1
Week - 2 January 15 – 19, 2024	LLM Architecture Libraries for LLMs	”Attention Is All You Need” (Vaswani et al., 2017) ”BERT: Pre-training of Deep Bidirectional Transformers” (Devlin et al., 2018)	Recap Quiz 1 Assignment 2
Week - 3 January 22 – 26, 2024	Guest Lecture on Generative AI Foundation Models	”Scaling Laws for Neural Language Models” (Kaplan et al., 2020)	Assignment 3
Week - 4 January 29 – February 2, 2024	Ethics of AI AI Security Fundamentals	”The Ethics of Artificial Intelligence” by Bostrom and Yud Kowalsky ”Concrete Problems in AI Safety” by Amodei et al	Recap Quiz 2 Assignment 4
Week - 5 February 5 – 9, 2024	AI Governance Frameworks AI Regulations *Kashmir Day‡: February 5, 2024*	”The EU AI Act: A Point of Comparison” by Veale and Borgesius ”Artificial Intelligence Governance: A Research Agenda” by Dafoe	Recap Quiz 3 Assignment 5
Week - 6 February 12 – 16, 2024	Vulnerabilities in LLMs Security Issues in LLMs	”Taxonomy of Attacks on Open-Source LLMs” by Hammond and Komatsu (2023) ”Red Teaming Language Models to Reduce Harms” by Ganguli et al. (2022)	Recap Quiz 4 Assignment 6
Week - 7 February 19 – 23, 2024	Advanced Security Testing for LLMs LLM Security Frameworks	”OWASP Top 10 for Large Language Model Applications” (2023) Selected technical documentation from GPT-Guard, LLM-Attacks	Recap Quiz 5 Assignment 7
Week - 8	Ethical Considerations in	”On the Dangers of Stochastic	Recap Quiz 6

February 26 – March 1, 2024	LLM Development Ethical Considerations in LLM Deployment	Parrots” by Bender et al. (2021) ”The Ethics of AI Ethics” by Hagendorff (2020)	Assignment 8
Week - 9 March 4 – 8, 2024	Guest Lecture on AI Governance Cloud Computing for AI	Selected case studies provided by guest lecturer ”Cloud Computing for Machine Learning and Cognitive Applications” by Kai Hwang	Reflection Discussion Assignment 9
Week - 10 March 11 – 15, 2025	LLM Security in AWS LLM Security in Google Cloud Platform (GCP) *1st Ramadan†: March 11, 2024*	”AWS Security Best Practices” - AWS Whitepaper ”Google Cloud Security Best Practices” - Google Cloud documentation	Recap Quiz 7 Assignment 10
Week – 11 March 18 – 22, 2024	LLM Security in Microsoft Azure Open-Source LLM Security *Conference Days: March 22 – 24, 2024 (No Classes)*	”Azure Security Best Practices and Patterns” - Microsoft Azure documentation ”The Hugging Face Transformers Book” - official documentation	Recap Quiz 8 Assignment 11
Week - 12 March 25 – 29, 2024	Hugging Face and LLM Security Enterprise Aspects of LLM Security *Last Day to Withdraw from Course(s): March 29, 2024*	”Security Best Practices for NLP Models" ”Enterprise AI Security: Challenges and Best Practices” - Industry whitepaper	Recap Quiz 9 Assignment 12
March 31, 2024	*21st Ramzan 1445 AH‡†*		
Week - 13 April 1 – 5, 2024	Future Trends in LLM Security Emerging Challenges in AI Security	The Landscape of AI Safety and Security Research” - Recent survey paper Selected recent publications from top AI security conferences	Recap Quiz 10
Week - 14 April 8 – 12, 2024	Final Project Presentations		

	Eid ul Fitr†‡: April 10 – 12, 2024		
Week - 15 April 15 – 19, 2024	Final Project Presentations		
Week – 16 April 22 – 26, 2024	Course Recap and Future Directions Career Paths in AI Security and Ethics *Last Day of Classes: April 26, 2024*		
April 27 - 29, 2024	*Reading Days*		
April 30, May 2 – 3 *&* *May 6 – 8, 2024*	*Labor Day‡: May 1, 2024* *Last Date to File Petition for Incomplete Grade: May 8, 2024*		
May 9 – 11, 2024			

Notes:

** The University reserves the right to correct typographical errors or to adjust the Academic Calendar at any time it deems necessary.*

† Subject to the sighting of the new moon.

‡ No Class(es).

§ University's Examination Policy is available in the Academic Policies folder on the Faculty/Staff Portal.

Attendance Policy

Students are expected to maintain 100% attendance in the courses at HU. However, all students must maintain class attendance per the attendance threshold document shared by the RO to deal with any

unforeseen situations at your end during the semester. If one cannot participate in any session, inform the instructor within 24 hours with a reason. Noncompliance will eventually lead to withdrawing/failing the student(s) from this course. Attendance will be marked manually in the class per the University's attendance policy.

Final Exam Policy

There will be no final exam but there will be a final project.

Academic Integrity

Each student in this course is expected to abide by the Habib University Student Honor Code of Academic Integrity. Any work submitted by a student in this course for academic credit will be the student's own work.

Scholastic dishonesty shall be considered a serious violation of these rules and regulations and is subject to strict disciplinary action as prescribed by Habib University regulations and policies. Scholastic dishonesty includes, but is not limited to, cheating on exams, plagiarism on assignments, and collusion.

a. Plagiarism: Plagiarism is the act of taking the work created by another person or entity and presenting it as one's own for the purpose of personal gain or of obtaining academic credit. As per University policy, plagiarism includes the submission of or incorporation of the work of others without acknowledging its provenance or giving due credit according to established academic practices. This includes the submission of material that has been appropriated,// bought, received as a gift, downloaded, or obtained by any other means. Students must not, unless they have been granted permission from all faculty members concerned, submit the same assignment or project for academic credit for different courses.
b. Cheating: The term cheating shall refer to the use of or obtaining of unauthorized information in order to obtain personal benefit or academic credit.
c. Collusion: Collusion is the act of providing unauthorized assistance to one or more person or of not taking the appropriate precautions against doing so.

All violations of academic integrity will also be immediately reported to the Student Conduct Office.

You are encouraged to study together and to discuss information and concepts covered in lecture and the sections with other students. You can give "consulting" help to or receive "consulting" help from such students. However, this permissible cooperation should never involve one student having possession of a copy of all or part of work done by someone else, in the form of an e-mail, an e-mail attachment file, a diskette, or a hard copy.

Should copying occur, the student who copied work from another student and the student who gave material to be copied will both be in violation of the Student Code of Conduct.

If you wish to use generative-AI tools to complete any of your assessments, you must first obtain permission from your course instructor. AI generated work will not be accepted in all classes or even all assessments. The instructor's permission is required. If the permission is granted, you should declare its use and properly cite the source of the generated content. Failing to identify AI written or assisted work is academic dishonesty and will be treated as any case of plagiarism by the university.

The principle for academic integrity is that your submissions must be substantially your own work and that any work that is not originally your thought must be identified and credited. If the use of AI tools is prohibited in the course, respect the rules and do not use these tools for assessments. The fundamental purpose of assessment is to learn, synthesize information and explain new connections and interpretations that arise from your secondary research. Be aware that unauthorized use of AI tools for assessments can result in a conduct case being filed. This can have serious consequences for your academic standing and future career opportunities.

During examinations, you must do your own work. Talking or discussion is not permitted during the examinations, nor may you compare papers, copy from others, or collaborate in any way. Any collaborative behavior during the examinations will result in failure of the exam, and may lead to failure of the course and University disciplinary action.

Penalty for violation of this Code can also be extended to include failure of the course and University disciplinary action.

Program Learning Outcomes (For Administrative Review)

Upon graduation, students will have the following abilities:

- PLO 1: Theoretical Computer Science: recall and apply foundational principles of computer science.
- PLO 2: Application Development: build software systems of varying complexity in light of fundamental computer science principles and any other constraints.
- PLO 3: Analysis and Design: perform technical analysis and design using core computing and mathematical knowledge.
- PLO 4: Systems: apply the knowledge of computing systems.

- PLO 5: Research and Exploration: develop expertise in and contribute to a given sub-field of computing by drawing upon a strong foundation in the fundamentals of computer science and mathematics to solve real-life problems.

- PLO 6: Problem Solving: identify and analyze problems and propose effective computing-based solutions.

- PLO 7: Practical Exposure: make effective use of current tools, technologies, and good industry practices.

- PLO 8: Responsible Citizenship: conduct their computing practice in a manner that is ethical and socially responsible and corresponds to their distinct sense of identity and service to the community.

- PLO 9: Self-Learning: continuously adapt their skills to the changes taking place around them.

- PLO 10: Design Thinking: apply design thinking principles to the design of a solution.

- PLO 11: Multi-disciplinarity: incorporate knowledge and input from multiple disciplines.

- PLO 12: Communication and Teamwork: communicate and function effectively as a member or a leader of a variety of teams.

Program Learning Outcomes (PLOs) mapped to Course Learning Outcomes (CLOs)					
	CLOs of the course are designed to cater following PLOs: **PLO 2: Application Development** **PLO 7: Practical Exposure** **PLO 8: Responsible Citizenship** **PLO 9: Self-learning**				
	Distribution of CLO weightages for each PLO				
	CLO 1	**CLO 2**	**CLO 3**	**CLO 4**	**CLO 5**
PLO 2					
PLO 7					
PLO 8					
PLO 9					

Mapping of Assessments to CLOs

Assignments	CLO #01	CLO #02	CLO #03	CLO #04	CLO #05
Assignment 1					
Assignment 2					
Assignment 3					
Assignment 4					
Assignment 5					
Assignment 6					
Assignment 7					
Assignment 8					
Assignment 9					
Assignment 10					
Assignment 11					
Assignment 12					
Final Project					

Quizzes	CLO #01	CLO #02	CLO #03	CLO #04	CLO #05
Quiz 1					
Quiz 2					
Quiz 3					
Quiz 4					
Quiz 5					
Quiz 6					
Quiz 7					
Quiz 8					
Quiz 9					
Quiz 10					

Recording Policy

Only asynchronous and synchronous online sessions will be conducted and recorded via MS Teams. Link to the recordings will be available to all students on Canvas Learning Management System.

Accommodations for Students with Disabilities

In compliance with the Habib University policy and equal access laws, I am available to discuss appropriate academic accommodations that may be required for student with disabilities. Requests for academic accommodations are to be made during the first two weeks of the semester, except for unusual circumstances, so arrangements can be made. Students are encouraged to register with the Office of Academic Performance to verify their eligibility for appropriate accommodations.

Inclusivity Statement

We understand that our members represent a rich variety of backgrounds and perspectives. Habib University is committed to providing an atmosphere for learning that respects diversity. While working together to build this community we ask all members to:

- share their unique experiences, values and beliefs
- be open to the views of others
- honor the uniqueness of their colleagues
- appreciate the opportunity that we have to learn from each other in this community
- value each other’s opinions and communicate in a respectful manner
- keep confidential discussions that the community has of a personal (or professional) nature
- use this opportunity together to discuss ways in which we can create an inclusive environment in this course and across the Habib community

Office Hours Policy

Every student enrolled in this course must meet individually with the course instructor during course office hours at least once during the semester. The first meeting should happen within the first five weeks of the semester but must occur before midterms. Any student who does not meet with the instructor may face a grade reduction or other penalties at the discretion of the instructor and will have an academic hold placed by the Registrar’s Office.