SHARDA UNIVERSITY
Beyond Boundaries

NAAC A+ ACCREDITED

Center for AI in Medicine, Imaging & Forensics (CAIMIF)

Organizing an

International Online Workshop on

BUILDING, TRUSTWORTHY GENERATIVE AI SYSTEMS: CONSISTENCY, RELIABILITY, EXPLAINABILITY, SAFETY AND TRUST (CREST)

in collaboration with **University of Maryland Baltimore County (UMBC), USA**

June 10, 2025
to
August 10, 2025
(Online)

Workshop Registration link: https://forms.gle/QtLWe4Sx3DgweV998

1. Workshop Design Criteria

This intensive and immersive hands-on workshop is designed for undergraduate students seeking to understand and develop trustworthy generative AI systems. The workshop takes an experiential learning approach that makes complex AI concepts accessible to students with varying levels of technical background. This workshop aims to bridge the gap between theory and practice, introducing undergraduates to the ethical and technical dimensions of AI while offering concrete tools and methodologies to evaluate and build generative AI systems that are consistent, reliable, explainable, safe, and trustworthy.

1.1 Workshop Description

This interactive hands-on workshop will introduce participants to the emerging field of NeuroSym-bolic AI approaches for building trustworthy generative AI systems. The workshop will start with a basic introduction to large language models (LLMs) and their capabilities and limitations. Understanding the in-depth working of generative AI models will lay the foundation for exploring various techniques to enhance their trustworthiness. This will be followed by creating real-life scenarios to implement methods that enhance consistency, reliability, explainability, and safety in generative AI systems. These exercises are targeted to empower practitioners to develop generative AI systems that align with human values and can be deployed in critical applications such as healthcare, legal systems, and mental health.

1.2 Schedule Overview

The workshop runs for 9 weeks from June 10 through August 10, with sessions organized as follows:

- Two lecture sessions per week (Tuesdays, Thursdays)
- One lab session every two weeks (Fridays)

2. Workshop Schedule

2.1 June - Month 1

Week 1

Tuesday – June 10

LECTURE 1: Introduction to Generative AI
Topics: LLM capabilities, challenges, and limitations
Key concepts: Foundation models, prompt engineering, hallucination issues

Thursday – June 12

LECTURE 2: CREST Framework Overview
Topics: NeuroSymbolic AI approaches, knowledge sources, framework components
Key concepts: Integration of neural and symbolic methods, knowledge representation

Friday - June 14

LAB 1: Exploring LLM Capabilities and CREST Framework
Activities:
- **Environment setup and configuration**
- **Prompt engineering practice with various LLMs**
- **Identifying common LLM issues (hallucination, inconsistency)**
- **Basic knowledge graph construction**
- **Attention visualization techniques**

Week 2

Tuesday - June 17

LECTURE 3: Consistency Challenges in Generative AI
Topics: Paraphrasing effects, knowledge-grounding techniques
Key concepts: Logical consistency, factual consistency, context sensitivity

Thursday - June 19

LECTURE 4: Improving Consistency
Topics: Self-consistency techniques, evaluation methods
Key concepts: Multiple sampling, consensus mechanisms, evaluation metrics

Week 3

Tuesday - June 24

LECTURE 5: Ensemble Approaches for Reliability
Topics: Knowledge-infused learning, shallow ensembling
Key concepts: Model combination strategies, uncertainty quantification

Thursday - June 26

LECTURE 6: Semi-Deep Ensembling
Topics: Domain knowledge integration, retrieval-augmentation
Key concepts: RAG systems, knowledge retrieval mechanisms

Friday - June 28

LAB 2: Consistency and Reliability Enhancement
Activities:
- Developing consistency measures
- Implementing self-consistency techniques
- Building model ensembles
- Testing with diverse inputs and adversarial examples
- Implementing basic retrieval augmentation

2.2 July - Month 2

Week 4

Tuesday - July 1

LECTURE 7: Explainable AI Fundamentals
Topics: User-level explainability, attention visualization
Key concepts: Interpretability methods, explanation formats

Thursday - July 3

LECTURE 8: Advanced Explainability Techniques
Topics: Evaluator pairing, knowledge retrievers, process knowledge
Key concepts: Explanation quality metrics, process transparency

Week 5

Tuesday - July 8

LECTURE 9: Safety Concepts in Generative AI
Topics: Red teaming approaches, contextual awareness
Key concepts: Threat modeling, safety evaluation, bias detection

Thursday - July 10

LECTURE 10: Process-Guided Safety
Topics: Safety constraints, abstention mechanisms
Key concepts: Knowledge boundaries, uncertainty handling

Friday - July 12

LAB 3: Explainability and Safety Implementation
Activities:
- Creating explainability visualizations
- Building evaluator pairs for explanation
- Developing safety evaluators
- Red teaming prompt design and testing
- Implementing guided safety constraints

Week 6

Tuesday - July 15

LECTURE 11: Knowledge Graphs and Symbolic Reasoning
Topics: Symbolic approaches, KG-based LLMs
Key concepts: Knowledge representation, logical inference

Thursday - July 17

LECTURE 12: Knowledge-Infused Learning
Topics: Mixture of experts, performance maximization
Key concepts: Specialized model combination, domain adaptation

Week 7

Tuesday - July 22

LECTURE 13: Healthcare AI Applications
Topics: Clinical knowledge integration, healthcare safety
Key concepts: Medical guidelines, clinical decision support

Thursday - July 24

LECTURE 14: Legal AI Applications
Topics: Legal reasoning, legal explainability
Key concepts: Citation mechanisms, precedent handling

Friday - July 26

LAB 4: Knowledge Integration and Domain Applications
Activities:
- Building knowledge representations
- Implementing symbolic reasoners
- Developing specialized model ensembles
- Creating clinical knowledge bases
- Building legal reasoning systems

2.3 August - Month 3

Week 8

Tuesday - July 29

LECTURE 15: Bias Detection in Generative AI
Topics: Attribution mechanisms, ethical considerations
Key concepts: Bias metrics, fairness evaluation

Thursday - July 31

LECTURE 16: Fairness in AI Systems
Topics: Content attribution, ethical guidelines
Key concepts: Fairness frameworks, ethical AI development

Week 9

Tuesday - August 5

LECTURE 17: Mental Health Applications
Topics: Clinical guidelines, mental health explainability
Key concepts: PHQ-9 integration, guideline adherence

Thursday - August 7

LECTURE 18: Project Development and Integration
Topics: System design principles, evaluation metrics
Key concepts: User-centered design, iterative development

Friday - August 9

LAB 5: Final Project Development and Presentations

Activities:

- **Team project presentations**
- **Peer and expert feedback**
- **Implementation of CREST framework components**
- **Ethical evaluation of developed systems**
- **Certificate ceremony and conclusion**

3. Session Type Summar

Type	Description	Total Sessions (40 Hrs)
Lectures	Theoretical foundations and concepts	18 (27 Hrs)
Labs	Hands-on implementation and practice	5 (10 Hrs)
Evaluation	Test	2 (3 Hrs)

4. Overall Workshop Objectives

- Understand the challenges and limitations of current generative AI systems
- Master techniques for enhancing consistency in generative AI outputs
- Learn methods for improving reliability through knowledge integration and ensembling
- Implement user-level explainability techniques for generative AI
- Develop safety-oriented approaches for responsible AI deployment
- Apply the CREST framework to real-world applications
- Create NeuroSymbolic AI systems that combine the strengths of neural and symbolic approaches

5. Workshop Structure

5.1 Hands-on Labs

Every lab session will include multiple practical exercises aimed at reinforcing the topics covered in the preceding lectures. The workshop will conclude with a team project where participants will design and implement a generative AI application using the CREST framework. These hands-on labs offer participants a technical understanding of generative AI and enable them to experience real-world challenges and solutions.

5.2 Lab Setup

- The majority of lab exercises will be done using pre-configured cloud-based notebooks
- Participants need to bring their own laptop (any modern laptop will suffice)
- Only a browser and internet connection are required for most exercises
- University computer labs will be available for students without personal laptops
- Teaching assistants will provide setup support before and during the workshop
- No advanced configuration or installation required

6. Additional Workshop Information

6.1 Pre-requisites

- Basic programming knowledge (preferably Python)
- Completion of an introductory computer science course
- No prior machine learning or AI experience required
- Interest in ethical and responsible technology development
- Willingness to engage in interdisciplinary thinking

6.2 Resources Provided

- Comprehensive workshop materials and code repositories
- Access to cloud-based computing resources for exercises
- Reference implementations of CREST framework components
- Reading list for continued learning
- Community forum for ongoing discussion and support

Workshop Fee:

Sharda University Students: **Free**
Sharda University Faculty:**1000/ INR**
Other Participants: **2000/ INR**
Foreign National: **$50**

Bank Details
for online payment

Bank Name :	ICICI Bank Ltd.
Branch Address :	Krishna Apra Royal Plaza, D-2, E(acb), Alpha-1, Greater Noida, Gautam Budh Nagar, UP- 201306
Account Holder Name :	Sharda University-Seminar
Account No. :	025405005815 (CURRENT AC)
IFSC Code :	ICIC0000254
SWIFT Code :	ICICINBBCTS
MICR Code :	110229037

Scan to pay

RESOURCE PERSON

Dr. Manas Gaur
Assistant Professor,
Computer Science &
Engineering Department, UMBC, USA

CONVENER
Prof. (Dr.) Ashok Kumar
Head, CAIMIF,
Sharda University

CO-ORDINATOR
Prof. (Dr.) Sanju Tiwari
Member, CAIMIF,
Sharda University

Sharda University
Plot No. 32, 34 - Knowledge Park 3, Greater Noida-201310, Delhi- NCR, India.