Systems for Generative AI [EECS 598 – Special Topics]

Lectures/Discussion: **TTh 10:30AM-12PM** | Projects/Makeups: F 2-3PM
Instructor: Mosharaf Chowdhury, SymbioticLab

Course Description

This class will introduce you to the key concepts and the state-of-the-art in practical, scalable, and fault-tolerant software systems for emerging Generative AI (GenAI) and encourage you to think about either building new tools or how to apply an existing one in your own research.

Since datacenters and cloud computing form the backbone of modern computing, we will start with an overview of the two. We will then take a deep dive into systems for the Generative AI landscape, focusing on different types of problems. Our topics will include: basics on generative models from a systems perspective; systems for GenAI lifecycle including pre-training, fine-tuning/alignment, grounding, and inference serving systems; etc. We will cover GenAI topics from top conferences that take a systems view to the relevant challenges.

Note that this course is **NOT** focused on AI methods. Instead, we will focus on how one can build software systems so that existing AI methods can be used in practice and new AI methods can emerge.

Logistics

Readings, Presentations, and Participation [50%]: The course will be paper reading-based (tentative schedule and reading list).

Every week, we'll read about 3–4 papers on average. Each student will write reviews for some of these papers to highlight their understanding of different aspects of the papers.

Each student, individually or as part of a group, will present at least one paper in class with a focus on identifying the key contributions of the paper.

Project [50%]: While readings will cover the "principles" behind systems for GenAI, a substantial course project will help students explore the "practical" side.

Each student will have to complete substantive work on an instructor-approved problem as part of a group and have original contribution. Surveys are not permitted as projects; instead, each project must contain a survey of background and related work.

Software graduate students will have to do a substantial course project, where they will design, build, and evaluate a new system.

Non-software students can consider an alternative strategy, where they can focus on verifying/measuring existing software systems instead of building a new system or fine-tuning a model on specific datasets and report their findings. This will introduce the students to the challenges in running GenAI systems in local laptops and in the cloud and applying these emerging tools in their own research.

Depending on the problem, the hands-on activities will be performed locally and in the cloud.

Prerequisites

Students are expected to have **good programming skills** and must have taken **at least one undergraduate-level systems-related course (from operating systems/EECS482,**

databases/EECS484, distributed systems/EECS491, and networking/EECS489). Having an undergraduate ML/AI course may be helpful, but not required or necessary.

Please contact the instructor with specific questions about workload and expectations. Undergraduates must receive explicit permission from the instructor to enroll.