CS 598 Fall 2026
Hardware and Software for AI

Instructor
Prof. Nishil Talati
nishil@illinois.edu
4116 Siebel Center for CS

TA
Yichao Yuan
yichaoy2@illinois.edu

Class Times & Location
Tuesdays and Thursdays
11:00 AM – 12:15 PM
1043 Sidney Lu Mech Engr Bldg

CRN: 69389

Course Website
https://courses.grainger.illinois.edu/cs598hsa/fa2026/

Other Websites Used in This Class
- Ed Discussion *(for announcements and online discussions)*
  https://edstem.org/us/courses/101107
- Gradescope *(for assignment handouts and submissions)*
  https://www.gradescope.com/courses/1345988

Course Description
This graduate seminar examines how modern AI workloads are optimized across the hardware and software stack. The course focuses on the systems principles behind efficient AI execution, including model and workload structure, kernels and compilers, runtime systems, memory management, distributed inference, agentic AI infrastructure, and specialized accelerator architectures. Rather than treating hardware and software as separate layers, the course emphasizes how algorithmic properties such as attention, state, sparsity, mixture-of-experts routing, diffusion denoising, and agent control flow interact with compute units, memory hierarchies, interconnects, and runtime scheduling.

The course is organized around research-paper readings and discussion together with a semester-long open-ended research project. Students will read influential and recent papers, present and critique them in class, and develop a research project that identifies a concrete bottleneck or opportunity in AI systems and evaluates a hardware, software, or hardware-software co-design solution.

Optional Textbook
No textbook.

Prerequisites

Students should be comfortable with computer systems and computer architecture concepts, including processors, memory hierarchies, parallelism, and performance analysis. Prior exposure to machine learning is helpful but not required; the course will provide some algorithmic background to reason about the systems behavior of LLMs, mixture-of-experts models, and AI agents. Programming experience in Python and familiarity with at least one systems programming environment (e.g., C/C++, CUDA, Triton, or a compiler/runtime framework) is recommended.

Format

- In-person graduate seminar with research-paper presentations and discussion.
- Student-led paper presentations, structured advocate/critic debates, and class-wide discussion.
- Semester-long open-ended research project in teams of four (five in rare cases) students.
- Guest lectures from leading industry researchers in AI systems and hardware design.
- Final Project Showdown organized as an open poster and demonstration session in the first-floor lobby of Siebel Center for Computer Science.

Paper Readings, Presentations, and Discussion

Most class meetings will center on two research paper. Students will work in groups of **three/four**, and three groups will take distinct roles for each set of papers. The presenting group will give a 35-minute technical presentation that explains the problem, background, key insight, design, methodology, results, and broader implications. A second group will serve as the paper's advocates and make the strongest possible case for acceptance: what is novel, what is technically strong, why the evaluation is convincing, and why the work matters. A third group will serve as critics and make the strongest case against acceptance: questionable assumptions, missing baselines, evaluation gaps, limitations, generality concerns, or alternative explanations. After the paper presentations, the advocate/critic debates will be conducted class-wide and will be moderated by the instructor.

The goal is not to determine a single 'correct' verdict, but to practice the kind of technical reasoning used in research groups and conference program committees. Roles will rotate across the semester so that students gain experience explaining a paper, defending it, challenging it, and participating in discussion. Students are expected to read the assigned paper before class even when their group is not presenting.

Semester-Long Research Project

The course project is an open-ended research project at the intersection of AI, computer systems, and computer architecture, completed individually or in teams of **four (five in rare cases) students**. Projects may study LLMs, agents, diffusion models, MoE models, training/post-training infrastructure, compilers, kernels, runtime systems, memory systems, distributed systems, CPUs, GPUs, TPUs, or emerging accelerators. Each project must begin with a clearly defined systems or hardware problem and identify at least one quantitative metric to optimize or understand, such as latency, throughput, goodput, HBM/DRAM traffic, utilization, communication volume, energy, memory footprint, or cost. Example directions include hardware modeling, kernel development on emerging hardware such as TPUs, optimizing PyTorch/Inductor or Triton-generated kernels,

runtime scheduling, KV-cache placement, CPU inference, communication-aware MoE execution, or energy-aware AI serving.

Teams will submit a 1-page proposal early in the semester describing the problem and motivation, research question or hypothesis, primary metric, proposed approach, evaluation methodology, baselines, feasibility, risks and mitigation plan, related work, and team responsibilities. Projects should establish a working baseline and initial characterization before the mid-semester review, then implement and evaluate the proposed design. The final deliverables will include a research-style report, project artifacts as appropriate, and a poster/demo for the Final Project Showdown. Negative results are acceptable when the project rigorously explains why an idea does not work, under what conditions it might work, and what was learned from the investigation.

For detailed timeline and instructions on the research project, please checkout the project guide document.

Grading Breakdown

Component	Weight
Semester-long research project	50%
Paper presentation	30%
Advocate/critic discussion roles	10%
Class participation and attendance	10%

Grading Policies

The project grade will reflect the quality of the problem formulation, technical depth, evidence and methodology, implementation or analysis effort, quality of the evaluation, clarity of the final report/poster, and the team's ability to explain what was learned. Project grades are not based solely on achieving a positive speedup or other target result.

The paper-presentation grade will reflect technical accuracy, clarity, organization, depth of analysis, quality of figures and explanation, preparedness for questions, and performance in the presentation/advocate/critic roles. Participation includes consistent attendance, evidence that assigned papers were read, substantive questions, and constructive engagement in class discussion.

Attendance and Participation

Because the course is discussion-driven, attendance is an essential part of the learning experience. Students should arrive prepared to discuss the assigned reading and to engage respectfully with presenters and classmates. The class participation will be graded based on the discussion in class and the online questions/criticism each student will submit during class. The attendance will be graded based on the best 12 out of 15 paper presentations. If an absence is necessary, students should notify the instructor when possible; repeated unexcused absences will reduce the participation/attendance component of the grade.

Collaboration

Project collaboration is expected within the approved project team. All team members must understand the full project and should contribute meaningfully to the design, implementation/analysis, evaluation, and presentation. Paper presentation groups may collaborate fully within their assigned role, but every student remains responsible for reading the assigned paper and participating independently in class discussion.

Use of Generative AI Tools

Generative AI tools may be used as research and programming aids unless otherwise stated for a particular activity. Students remain responsible for the technical correctness, originality, and attribution of all submitted work. For project reports and presentations, students should disclose substantive use of generative AI tools when those tools materially contributed to code, analysis, writing, or figure generation.

Academic Integrity

The University of Illinois Urbana-Champaign Student Code should also be considered as a part of this syllabus. Students should pay particular attention to Article 1, Part 4: Academic Integrity. Read the Code at the following URL: https://studentcode.illinois.edu/

Every student is expected to review and abide by the Academic Integrity Policy as defined in the Student Code (https://studentcode.illinois.edu/article1/part4/1-401/). Academic dishonesty will result in a sanction proportionate to the severity of the infraction, with possible sanctions described in 1-404 of the Student Code (https://studentcode.illinois.edu/article1/part4/1-404/). As a student it is your responsibility to refrain from infractions of academic integrity and from conduct that aids others in such infractions. A short guide to academic integrity issues may be found on the provost's website (https://provost.illinois.edu/policies/policies/academic-integrity/students-quick-reference-guide-to-academic-integrity/). Ignorance of these policies is not an excuse for any academic dishonesty. It is your responsibility to read this policy to avoid any misunderstanding. Do not hesitate to ask the instructor(s) if you are ever in doubt about what constitutes plagiarism, cheating, or any other breach of academic integrity.

While this is not comprehensive, you may not copy any answers or code from online resources, from ChatGPT, or from other students. If you must include a quote or a figure from an external source (e.g., for a presentation; not allowed for homework or exams), cite the source to avoid plagiarism.

Mental Health

Diminished mental health, including significant stress, mood changes, excessive worry, substance/alcohol abuse, or problems with eating and/or sleeping can interfere with optimal academic performance, social development, and emotional well-being. The University of Illinois offers a variety of confidential services including individual and group counseling, crisis intervention, psychiatric services, and specialized screenings at no additional cost. If you or someone you know experiences any of the above mental health concerns, it is strongly encouraged

to contact or visit any of the university's resources provided below. Getting help is a smart and courageous thing to do – for yourself and for those who care about you.

- Counseling Center: +1 (217) 333-3704, 610 East John Street Champaign, IL 61820
- McKinley Health Center: +1 (217) 333-2700, 1109 South Lincoln Avenue, Urbana, IL 61801
- University Wellness Center: https://wellness.illinois.edu/

CS CARES, CS Values, and Code of Conduct

All members of the Illinois Computer Science department – faculty, staff, and students – are expected to adhere to the CS Values and Code of Conduct (https://cs.illinois.edu/about/values). The CS CARES Committee (https://cs.illinois.edu/about/cs-cares/) is available to serve as a resource to help people who are concerned about or experience a potential violation of the code. If you experience such issues, please contact the CS CARES Committee (https://cs.illinois.edu/about/cs-cares/contact). The instructor of this course is also available for issues related to this class.

Anti-Racism and Inclusivity Statement

The Grainger College of Engineering is committed to the creation of an anti-racist, inclusive community that welcomes diversity along a number of dimensions, including, but not limited to, race, ethnicity and national origins, gender and gender identity, sexuality, disability status, class, age, or religious beliefs. The college recognizes that we are learning together in the midst of the Black Lives Matter movement, that Black, Hispanic, and Indigenous voices and contributions have largely either been excluded from, or not recognized in, science and engineering, and that both overt racism and micro-aggressions threaten the well-being of our students and our university community.

The effectiveness of this course is dependent upon each of us to create a safe and encouraging learning environment that allows for the open exchange of ideas while also ensuring equitable opportunities and respect for all of us. Everyone is expected to help establish and maintain an environment where students, staff, and faculty can contribute without fear of personal ridicule, or intolerant or offensive language. If you witness or experience racism, discrimination, micro-aggressions, or other offensive behavior, you are encouraged to bring this to the attention of the course director if you feel comfortable. You can also report these behaviors to the Bias Assessment and Response Team (BART; https://bart.illinois.edu/). Based on your report, BART members will follow up and reach out to students to make sure they have the support they need to be healthy and safe. If the reported behavior also violates university policy, staff in the Office for Student Conflict Resolution may respond as well and will take appropriate action.

Disability-Related Accommodations

To obtain disability-related academic adjustments and/or auxiliary aids, students with disabilities must contact the course instructor and the Disability Resources and Educational Services (DRES) as soon as possible. To contact DRES, you may visit 1207 S. Oak St., Champaign, call +1 (217) 333-4603, e-mail disability@illinois.edu, or go to https://www.disability.illinois.edu/. If you are concerned you have a disability-related condition that is impacting your academic progress, there are academic screening appointments available that can help diagnosis a previously undiagnosed

disability. You may access these by visiting the DRES website and selecting "Request an Academic Screening" at the bottom of the page.

Family Educational Rights and Privacy Act (FERPA)

Any student who has suppressed their directory information pursuant to the Family Educational Rights and Privacy Act (FERPA) should self-identify to the instructor to ensure protection of the privacy of their attendance in this course. See https://registrar.illinois.edu/academic-records/ferpa/ for more information on FERPA.

Religious Observances

Illinois law requires the University to reasonably accommodate its students' religious beliefs, observances, and practices in regard to admissions, class attendance, and the scheduling of examinations and work requirements. You should examine this syllabus at the beginning of the semester for potential conflicts between course deadlines and any of your religious observances. If a conflict exists, you should notify your instructor of the conflict and follow the procedure at https://odos.illinois.edu/community-of-care/resources/students/religious-observances/ to request appropriate accommodations. This should be done in the first two weeks of classes.

Sexual Misconduct Reporting Obligation

The University of Illinois is committed to combating sexual misconduct. Faculty and staff members are required to report any instances of sexual misconduct to the university's Title IX Office. In turn, an individual with the Title IX Office will provide information about rights and options, including accommodations, support services, the campus disciplinary process, and law enforcement options.

A list of the designated university employees who, as counselors, confidential advisors, and medical professionals, do not have this reporting responsibility and can maintain confidentiality, can be found at https://wecare.illinois.edu/resources/students/#confidential

Other information about resources and reporting is available at https://wecare.illinois.edu/