CS59200-DMAI: Data Management for AI

Instructor: Chunwei Liu
Email: chunwei@purdue.edu
Office: LWSN 2142D
Time: TBD
Location: TBD

Course Description:

This graduate seminar examines the evolving frontier at the intersection of data management and artificial intelligence. In recent years, large language models (LLMs) and related AI advances have transformed once-thorny tasks such as information extraction, schema matching, and data integration into problems that can be addressed with just a few well-crafted prompts. Nevertheless, significant hurdles remain in turning these prototypes into production-grade, cost-efficient, and high-quality data pipelines.

This seminar will explore how classic database techniques—such as declarative interfaces, cost models, indexes, query planning, space and computation optimization, and workload-aware scheduling—are being reimagined for LLM training and inference. In parallel, we will examine the increasing role of AI systems, including vector databases and LLMs, in automating semantic operations within modern data management. Throughout, we'll investigate the symbiotic relationship between advances in machine learning and foundational database technologies: how machine learning can solve classical database challenges, and how robust data systems can power scalable, reliable AI solutions. Ultimately, the goal is to achieve database-class performance and reliability for unstructured data, including text, images, PDFs, and more, unlocking analytics from the limitations of traditional tabular approaches and enabling the next generation of intelligent applications.

Format:

This class will be structured as a seminar with a semester-long project component. The students will present, review and discuss papers. As groups, they will also conduct a research project over the course of the semester. In addition, there will be several guest lectures by researchers and professional experts in the area.

- Weekly discussions of cutting-edge research papers
- Invited talks from industry and academia on real-world LLM+data prototypes/systems
- Student presentations on projects/papers with emerging ideas and/or open problems
- A collaborative, semester-long research project

Prerequisites:

This course is intended for students interested in the intersection of databases, systems, machine learning, and AI. While prior coursework in databases (e.g., CS44800, CS54100, or CS54200) is beneficial, it is not required—essential database background material will be covered as needed. Please note that this is a database systems course and will not focus on core AI or machine learning basics. However, as the class will frequently refer to and utilize AI techniques and technologies, students are expected to have some familiarity and hands-on experience with these areas to fully participate and succeed.

Learning Outcomes:

Participants will leave with a concise, system-level toolkit for designing faster, cheaper, and more reliable generative AI applications, as well as a deep understanding of how data infrastructure and foundation models can co-evolve to power the next generation of analytics.

Grading Criteria*:

- Paper reviews (Guidance TBA): 5%
- Paper presentation evaluation (Guidance TBA): 15%
- Project milestone evaluation: 25% (Suggested or self-proposed projects)
- Participation: 5%
- Midterm exam: 25%
- Final Project evaluation: 25%

Syllabus (TBA)*:

References (TBD):

Palimpzest, Lotus, DocETL, EvaDB, DSPy, Evaporate, DB-BERT, GraphRAG, Fondation Databases, ...

*Disclaimer. This syllabus and grading criteria are subject to change.