Fall 2025 - CMPT 984 G100

Special Topics in Databases, Data Mining, Computational Biology (3)

Class Number: 5577

Sep 3 – Dec 2, 2025: Tue, 2:30–5:20 p.m.

Instructor: Ke Wang

Course details

This is a seminar-style special topics course on trustworthy AI and risks and robustness of LLMs. The purpose is to promote the awareness of safety and security risks of machine learning (ML) and LLM technology, and to understand technical solutions to these problems. Teaching materials are a collection of slides and a reading list consisting of survey papers and research papers on trustworthy AI, LLMs, and safety alignments for LLMs. All materials are hosted at the course website at sfu.canvas/Files. Working knowledge of machine learning is assumed.

Topics

Trustworthy AI (Explainable, interpretable, fair, robust, transparent, safe and secure), attacks on ML and defenses, attacks on LLMs and safety alignment of LLMs, multimodal LLMs, attacks, and defenses.

Schedule

This graduate course will be run in three formats: lecture presentation by instructor, paper presentation by students, and project presentation by students. The schedule is as follows

- Lecture weeks (4 weeks): Sept 9, Sept 16, Sept 23, Oct 7
- Paper presentation weeks (4 weeks): Oct 14, Oct 21, Nov 4, Nov 18,
- Project presentation weeks (2 weeks): Nov 25, Dec 2.
- Proposal presentation weeks: Oct 28

Grading

Class participation (attendance and discussion):10%

Paper presentation: 30%

Project proposal: 10%

Final project report: 30%

Project presentation: 20%