CMSC 848N Generative AI Agents

Computer Science

Furong Huang

https://furong-huang.com/

Tu/Th 12:30pm – 1:45pm
Sections 0101 56, PJ01 24

Course Introduction

- Syllabus, Topics and Prerequisite

Format

Lectures + Homework + **Course Project** + Mid Report/Office Hour + Final Presentation/Report

Tu/Th 12:30pm – 1:45 pm

- **Lectures (10% In-class impromptu quiz)**: Sep 2 – Nov 27
  - Recess: Oct 14 Fall Break, Nov 27 Thanksgiving
  - **TA led Q&A**: Sep 30
- **Final presentations (20%)**: Dec 2, Dec 4, Dec 9, Dec 11

Midterm Report (30%): Oct 15 (30%)

Final Report (20%): Dec 12 (40%)
Homework (20%): Sep 30, Oct 30, Nov 30

Late Bank: everything we ask for with a due date (expect for in-person exams) have a **total of 72 hours** late bank. Use wisely.

Lectures

- Slides, reading material (research papers) + anything else you need to learn to digest (self-learning)
- Interactive lecture, instructor-led presentation + Q&A + breakout discussion
- Find partner(s) for course projects

What do students expect to gain?

- Learn AI/ML frontier
- Learn reading papers effectively
- Practice presentation skills
- Establish teamwork
- Gain research skills through projects
  - Search, read & digest relevant literature
  - Reproduce baselines
  - Produce novel ideas
  - Implement your ideas
  - Run experiments
  - Collect & present your results
  - Work with fellow researchers, mentors (instructor/TAs)
  - Draft paper and slides

What are some expectations from the Instructor?

- You show up for yourself
  - Read papers, attend lectures, start your project early
  - Be social, seek information, find project partners, ask questions/suggestions, search solutions
  - Be critical, creative and open minded
    - Be critical about existing papers
    - Create your novel solutions
    - Be prepared to start over if things don’t work as expected, that is what research is about, learn from failures, big heart!
- You show up for your partners
  - Be collaborative in projects
  - Be responsive, be responsible, be punctual
  - Be supportive, patient, communicative, friendly

Course Description

- Foundations and frontiers of Generative AI Agents
- Theory and practice
- SoTA research in
  - LLMs;
  - RL;
  - Alignment;
  - Reasoning models;
  - Self-improvement;
  - Agent safety..

Learning Outcomes

By the end of this course, students will:

- Understand key architectures and algorithms for generative AI agents.
- Critically analyze and reproduce state-of-the-art methods.
- Design, implement, and evaluate novel AI agent systems.
- Conduct research suitable for submission to top-tier AI/ML venues.

Tentative Schedule – Part 1

- W1 [Sep 2, Sep 4] Introduction to AI Agents and RL – Definitions, taxonomies, historical context.
- W2 [Sep 9, Sep11] Foundations of RLHF – DPO, GRPO, and bilevel optimization (PARL)
- W3 [Sep 16, Sep 18] Alignment Challenges -MaxMin-RLHF, Test-time Alignment – Transfer-Q, GenARM, Collab
- W4[Sep 23, Sep 25] Reasoning Models – Chain-of-thought, Monte Carlo Tree Search, ThinkLite-VL
- W5[Sep 30, Oct 2] TA-led Q&A, Self-improvement – EnsemW2S, SoTA with Less, MORSE-500

Tentative Schedule – Part 2

- W6 [Oct 7, Oct 9] Agentive Workflows – Design patterns, communication graphs, role optimization
- W7 [Oct 16] Overflow
- W8 [Oct 21, Oct 23] Web Agents – Architectures, capabilities, vulnerabilities
- W9 [Oct 28, Oct 30] Code Agents – Code generation, debugging, and security considerations
- W10 [Nov 4, Nov 6] Tool-Use Agents – Tool integration, orchestration frameworks

Tentative Schedule – Part 3

- W11 [Nov 11, Nov 13] World Models – For web, robotics and simulation-based agents
- W12 [Nov 18, Nov 20] Safety and Robustness – Jailbreak, poisoning, and agentic defenses
- W13 [Nov 25, Nov 27] AI-Generated Content Detection – watermarking, and detectors.

Prerequisite – Part 1

- Basic machine learning concepts
  - Supervised/unsupervised/reinforcement learning
  - Classification, Regression, Cross validation, Overfitting, Generalization
  - Deep neural networks
  - See math4ml
- Basic calculus and linear algebra
  - Compute (by hand) gradients of multivariate functions
  - Conceptualize dot products and matrix multiplications as projections
  - Solve multivariate equations using, etc, matrix inversion, etc.
  - Understand basic matrix factorization
  - See linear algebra review, and advanced

Prerequisite – Part 2

- Basic optimization
  - Use techniques of Lagrange multipliers for constrained optimization problems
  - Understand and be able to use convexity
  - See convex analysis review, optimization review
- Basic probability and statistics
  - Understand: random variables, expectations and variance
  - Use chain rule, marginalization rule and Bayes' rule
  - Make use of conditional independence, and understand "explaining away"
  - Compute maximum likelihood solutions for Bernoulli and Gaussian distributions
  - See probability review

Action Items

- Instructors upload a catalog of course projects
  - Highly recommend choosing one of the list
  - In very rare cases, if you would like to work on your own project, you must convince me that it is highly relevant to our course
- Students sign up for course projects, book meetings with tentative project partners. Theoretically, no bigger than a group size of 3 (3 is allowed).
- TAs set up slack for asynchronous office hours

Ice Break

Instructors

Instructor:
Furong Huang

TA:
Minghui Liu

TA:
Ho Sy Tuyen

FURONG'S LAB @ UMD

CMSC 848N Generative AI Agents

https://furong-huang.com/

furongh@umd.edu

DEPARTMENT OF COMPUTER SCIENCE

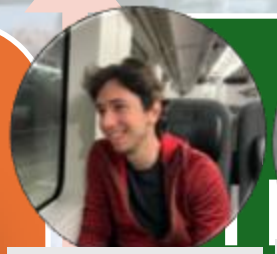
CENTER FOR MACHINE LEARNING

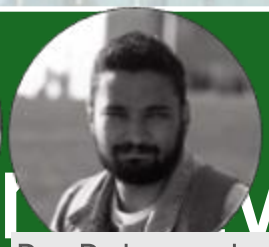
Current Students

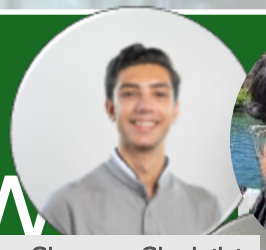
Em

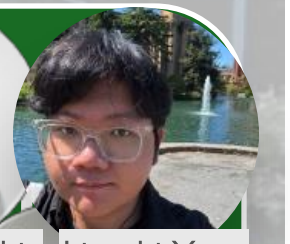

Candidate Ruijie Zheng


Seungjae Lee

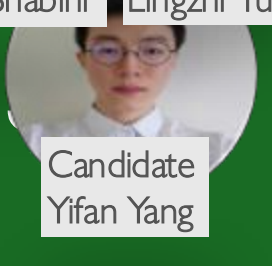

Michael-Andrei


Pan Pathmanathan


Shayan Shabihi


Lingzhi Yuan

Agent


Candidate Yifan Yang


Candidate Chenghao Deng


Postdo
Zikui C


Candidate Mucong Ding


Candidate Souradip Charkraborty


Andrew Mendez

Reasonin


Minghui Liu


Weize Liu


Aakriti Agrawal


Yongyuan Liang


Candidate Xiyao Wang

Multimodal


Ho Sy Tuyen

Our Solutions

IVE [CoRL'25]
FLARE [CoRL'25]
PEnGUiN [RLC'25]
TraceVLA [ICLR'25]
World Models w Hints of LLMs [NAACL'25]
Make-An-Agent [NeurIPS'24]
PRISE [ICML'24]
Premier-TACO [ICML'24]
COPlanner [ICLR'24]
DrM [ICLR'24]

Model Tampering Attacks [TMLR'25]
Immune [CVPR'25]
MergeME [NAACL'25]
PoisonedParrot [NAACL'25]
GFairHint [TKDD'25]
Safe MARL [ICRA'25]
Is Poisoning a Real Threat to DPO? [AAAI'25]
Watermark prevent copyright gen [AAAI'25]
FACT or Fiction [NeurIPS'24]
Shadowcast [NeurIPS'24]
AutoDAN [COLM'24]
Possibilities of AI-Generated Detection [ICML'24]
TrustLLM [ICML'24]
Beyond Worst-case Attacks [ICLR'24]

GenARM [ICLR'25]
GenFlowRL [ICCV'25]
Collab [ICLR'25]
CSRec [SIGIR'25]
LLM and Causal Infer in Collaboration [NAACL'25]
Easy2Hard-Bench [NeurIPS'24]
Transfer-Q-star [NeurIPS'24]
MaxMin-RLHF [ICML'24]
PARL [ICLR'24]
SAFLEX [ICLR'24]
C-Disentanglement [NeurIPS'23]

GenFlowRL [ICCV'25]
VisVM [ICCV'25]
Zero-Shot Encoder Graft [ICCV'25]
VLM Unlearning Bench via Fictious Data [ICLR'25]
SIMA [NAACL'25]
AUTOHALLUSION [EMNLP'24]
Mementos [ACL'24]
WAVES [ICML'24]
HallusionBench [CVPR'24]
More Context, Less Distraction [ICLR'24]