

AI Security: Course Logistics

Sangdon Park

Assistant Professor

Machine Learning Lab

Graduate School of AI (**GSAI**)

Computer Science and Engineering (**CSE**)

POSTECH

POSTECH



TA

- Sechan Lee (GSAI)



Textbook

- No Official Textbook

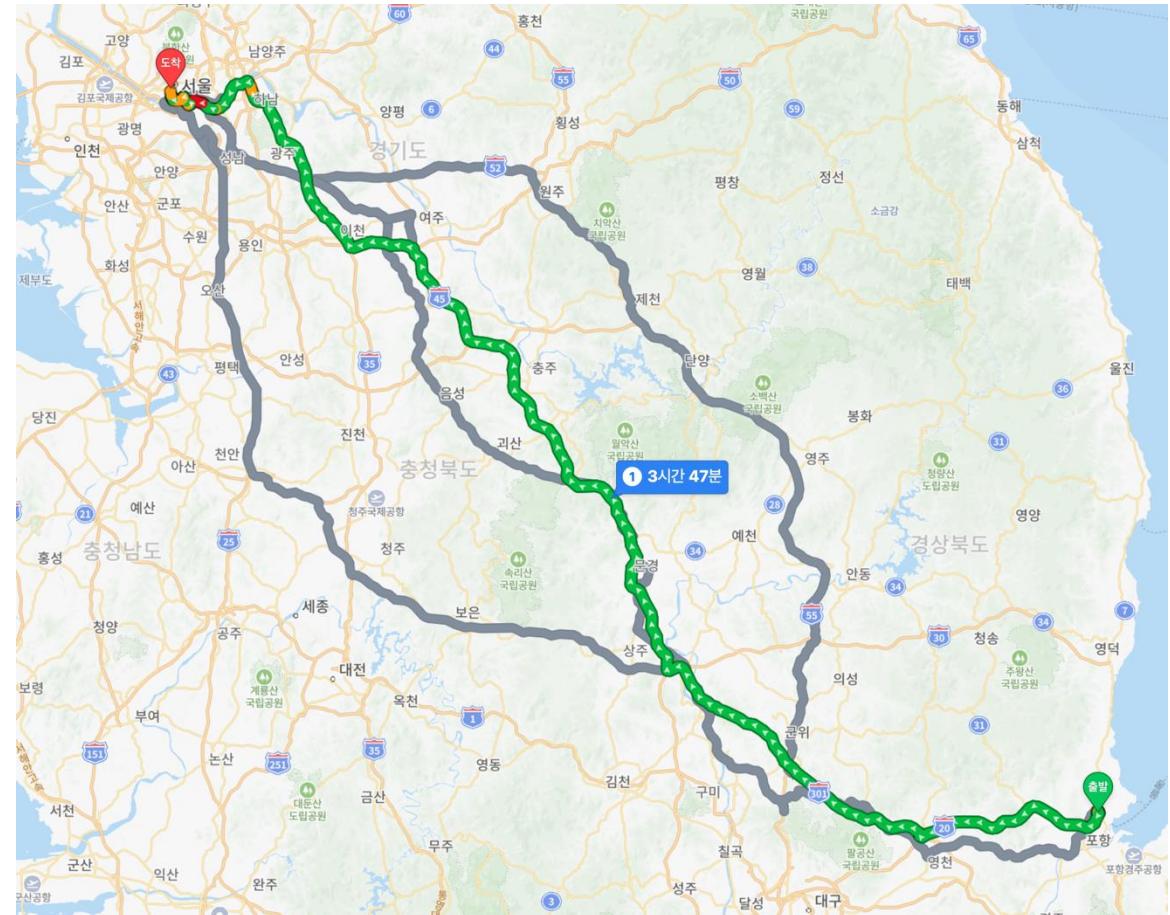
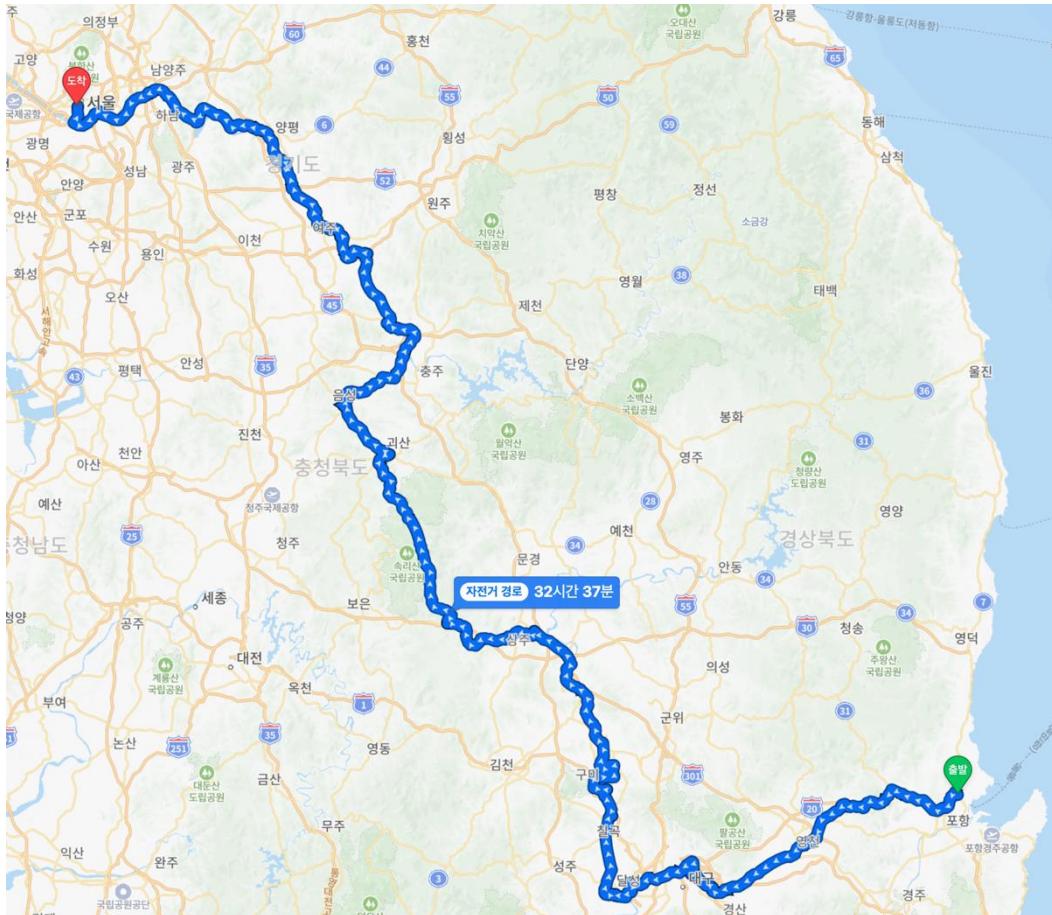
Grading

- Assignment / Presentation (**80%**)
 - HW1 (**10%**)
 - HW2 (**20%**)
 - Final Project (**50%**)
- Participation (**20%**)
 - 5% for each (question, answer) pair
- Note
 - Grad students – recommend to audit.

GenAI Usage Policy

Should We Use Generative AI?

- Should We Use A Car?



GenAI Usage Policy

Should We Use Generative AI?



Andrew Ng

DeepLearning.AI

THE BATCH

September 3, 2025

What Matters in AI Right Now

[Subscribe](#) [Submit a tip](#)

Dear friends,

There is significant unmet demand for developers who understand AI. At the same time, because most universities have not yet adapted their curricula to the new reality of programming jobs being much more productive with AI tools, there is also an uptick in unemployment of recent CS graduates.

When I [interview](#) AI engineers — people skilled at building AI applications — I look for people who can:

- Use AI assistance to rapidly engineer software systems
- Use AI building blocks like prompting, RAG, evals, agentic workflows, and machine learning to build applications
- Prototype and iterate rapidly

GenAI Usage Policy

Should We Use Generative AI?



Andrew Ng



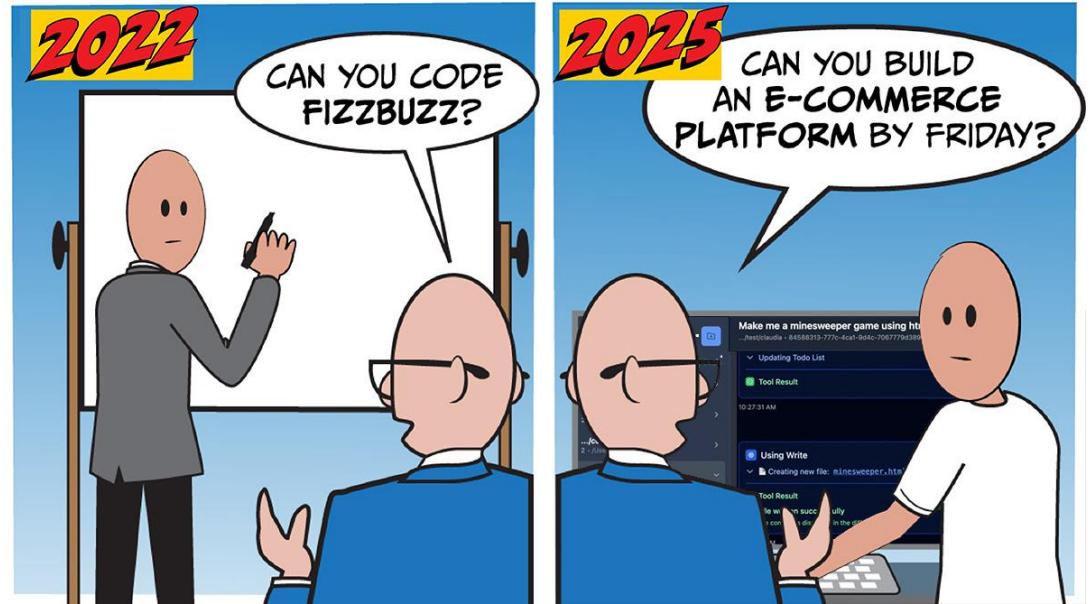
Punchcard Programmer

GenAI Usage Policy

Should We Use Generative AI?



Andrew Ng



GenAI Usage Policy

- Coding Agents – OK but carefully verify results
- QA Agents – OK but carefully verify results
- Whatever Agents – OK but carefully verify results
- Feel free to use
 - However, intensitvely use it as your peers will use it and I'll use a relative grading system.

Homeworks

- HW1
 - Choose an interesting victim **non-generative** model
 - Image classifiers, regressors, object detectors, ...
 - Define an attack goal (~ = a security goal)
 - Implement an attacker to achieve the attack goal.
 - Present your method in about 20 mins
 - Persuade me that your selected victim model, attack goal, and method are interesting.
 - You can also share your failure experience.
 - You should answer my questions **on details**.
 - Probably we may not find novel attack goals or attack methods; that's fine.

Homeworks

- HW2
 - Choose an interesting victim **generative** model
 - LLMs, LRM_s, LCM_s, diffusion models, VLM_s, VLAs
 - Define an attack goal (~ = a security goal)
 - Implement an **existing** attacker to achieve the attack goal.
 - Present your method in about 20 mins
 - Persuade me that your selected victim model, attack goal, and method are interesting.
 - You can share your failure experience.
 - You should answer my questions **on details**.

Final Project

- Use **OpenClaw** for your victim AI model
- Define an attack goal (~ = a security goal)
- **Propose** an attacker to achieve the attack goal
 - Justify why your attacker is novel via table comparison
 - Again, you can brainstorm with GenAI on novel ideas.
- Present your method in about 30 mins
 - You can share your failure experience **but has a little penalty**.
 - You should answer my questions **on details**.
 - Your method is considered to be novel if ChatGPT says so.
- (optional) Write a paper in the case of undergrads

Q&A

Welcome any feedback on the HW/Project style