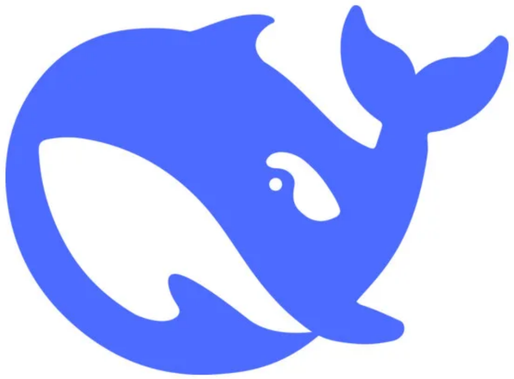
Introduction & Administrivia

EECE695D: Efficient ML Systems

Spring 2025

What happened?

$560B = 2.7 * R&D budget of Korean government



Semiconductor

NVIDIA -17%

tsmc -13%

arm -10%

ASML -7%

APPLIED MATERIALS -7%

IT Hardware

ARISTA -22%

DELL -9%

Data Centers

DIGITAL REALTY -9%

EQUINIX -4%

Electricity

VISTRA -28%

TALEN ENERGY -22%

Constellation -21%

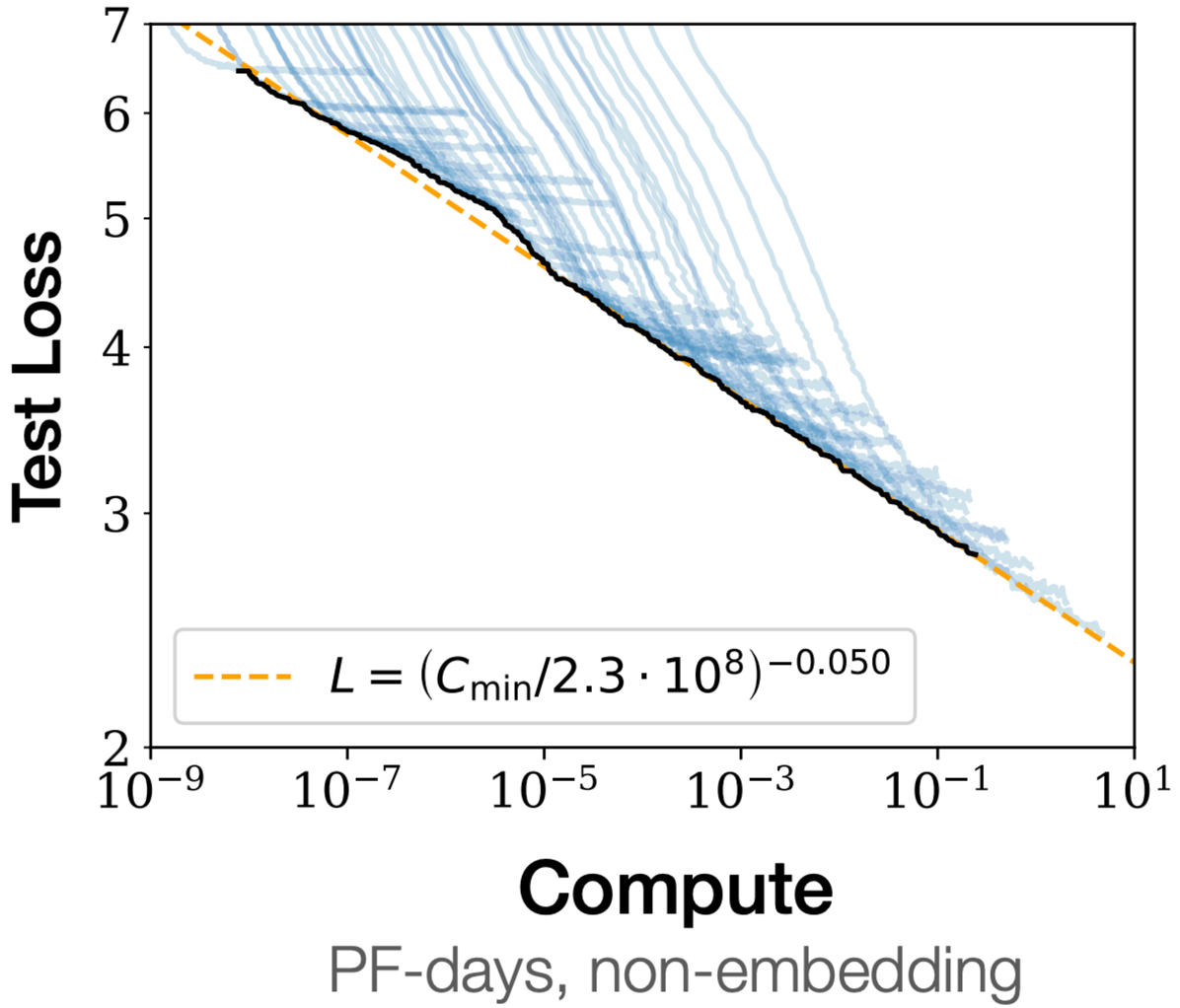
Why was DeepSeek so disruptive?

- **New Capability?** No!

- **Cheaper Cost?** Yes ;)
  - $6M for LLM pre-training
    ($100M for GPT-4, reportedly)
  - 2,000 H800 GPUs
    (6,000—10,000 H100 for GPT-o1, reportedly)

That means...

- Given the same budget, we can
  - use larger models
  - train longer
  - use more data
  - … and get better performance!



Kaplan et al., "Scaling Laws for Neural Language Models," arXiv 2020

Also, the inference cost

- On the inference side, cheaper inference cost means:
  - Cheaper LLM subscription
  - Smarter on-device AI
  - Better reasoning
  - Better generation quality

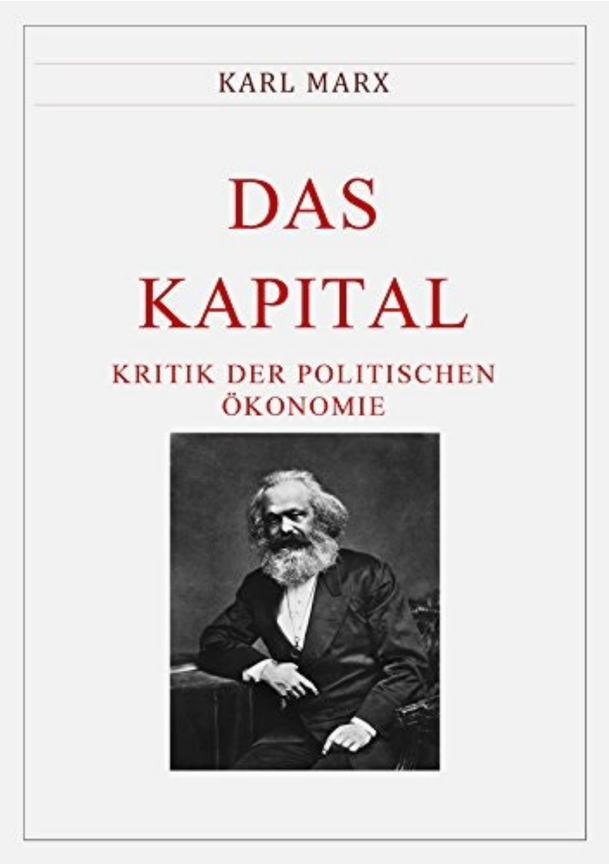
1x Compute

32x Compute

https://openai.com/index/video-generation-models-as-world-simulators/

Humanitarian Issues

- Democratization
- Environment & Risk



About this course

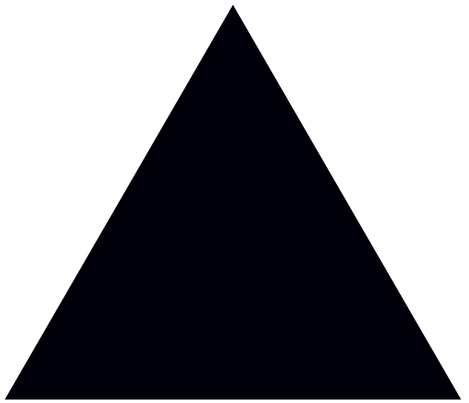
Coverage

- Efficient AI, from the algorithmic side

- **Goal.** Help you confidently answer…
  - How much computation & memory will my model need?
  - How can we make my model run faster?
  - How can we make my model be lighter?
  - How can I reduce the training cost?

Ultimately...

- Ultimately, we want to understand the grand tradeoff:
  - Prediction Quality
  - Training Cost
  - Inference Cost



e.g., linear model is cheaper to train(?), cheaper to predict, but achieves lower performance than deep, overparameterized neural networks

- But our knowledge is not quite there yet ;(
  - So let's be a bit more practical and applied!

Schedule

Phase 1. Lecture Only

- **W1.** Warm-Up
- **W2.** Sparsity
- **W3.** Quantization
- **W4.** NAS & KD
- **W5.** Efficient Training & Tuning
- **W6.** Adaptation
- **W7.** (No class; EU-KR AI meetings)
- **W8.** Parallelism

Phase 2. Lecture + Student Presentation

- **W9.** Data Efficiency
- **W10.** LLM Compression
- **W11.** Long-Context LLMs
- **W12.** Low-Precision Training
- **W13.** Test-time Scaling
- **W14.** Efficient Diffusion Model
- **W15.** Efficient Neural Rendering
- **W16.** Final Presentation

Prerequisites

- **Necessary**
  - Familiar with deep learning
    - Know how to code in Python
    - Experiences with PyTorch / TensorFlow / Jax
  - Have read 10+ academic papers
- **Recommended**
  - Knows how CPU / GPU / memory works

Administrivia

Team

- **Instructor.** Jaeho Lee jaeho.lee@postech.ac.kr
  - Assistant Professor @ POSTECH EE
  - Visiting Researcher @ Google
  - Lectures
- **Teaching Assistant.** Hyunjong Ok hyunjong.ok@postech.ac.kr
  - M.S./Ph.D. Candidate @ POSTECH GSAI
    - Much expertise on Language & Speech
  - Assignments & Attendance (but please cc me as well)

Hours & Location

- **Lectures**
  - PIAI 122
  - MW 11:00—12:15
- **Office Hours**
  - Terarosa Coffee
  - W 17:00—18:00 + by appointment
  - Tip. Think of it as a friendly coffee chat ☕

Resources

- **Textbook.** None
    - Slides uploaded at jaeho-lee.github.io
        - some references there as well

Grading

• Attendance	10%
• In-class Presentation	30%
• Project — Proposal	20%
• Project — Final	40%
	100%

- **Note.** You get an F, if you (1) miss 3+ classes
  (2) use chatbots for assignments
  (3) cheat

In-Class Presentation

- From W9—W15, we do
    - **Mon.** Lecture
    - **Wed.** Student Presentation
- By W2. I'll give you a paper list.
    - 3 papers for each week
    - By W3, You sign up as a presenter for the paper
        - Maximum 2 per paper

In-Class Presentation

- At your presentation week, give a 20min talk, 5min Q&A.

- **Rubrics**
  - How clearly did you describe the research question?
  - How well did you describe the proposed solution?
  - How well did you identify the limitations
    (esp. the ones not mentioned in the paper)
  - How much did you engage in others' presentations?

Project

- You'll prepare a submission to the ICLR blog post track, for ICLR 2026
  - https://iclr-blogposts.github.io/2025/about/

- **Idea**

1. Reviews past work and summarize the outcomes, develop new intuitions, or highlight some shortcomings.
2. Presents novel perspectives or interpretations of existing machine learning concepts or techniques.
3. Discusses important issues in machine learning, such as reproducibility, from a novel perspective.
4. Analyzes the societal implications of recent advancements in machine learning and AI.
5. Showcases cool research ideas that you tried but did not work out.

  - Plus, make it about Efficient ML

Project

- **Cool Examples**
  - A Deeper Look at Zero-Cost Proxies for Lightweight NAS https://iclr-blog-track.github.io/2022/03/25/zero-cost-proxies/
  - Unraveling the Impact of Training Samples https://iclr-blogposts.github.io/2024/blog/unraveling-the-impact-of-training-samples/
  - Understanding In-Context Learning in Transformers https://iclr-blogposts.github.io/2024/blog/understanding-icl/

Project

- Max 3 people per group.
- **Proposal (W7)**
    - Write a one-page description of what you'll do. LaTeXed.
- **Final Poster Session (W16)**
    - By W15, submit your blog post via PLMS
    - By W15, Prepare a (small) poster, in pdf
    - I'll print 'em out and get some 🍕
    - We do a small poster session of our own

Project

- **Rubrics**
    - Clarity
    - Soundness
    - Originality
    - TA’s assessments
    - Peer assessments

That’s it for today 🙌